

Probabilistic Integration: A Role for Statisticians in Numerical Analysis?

François-Xavier Briol¹, Chris. J. Oates^{2,3}, Mark Girolami^{1,4},
Michael A. Osborne⁵ and Dino Sejdinovic⁶

¹Department of Statistics, University of Warwick

²School of Mathematical and Physical Sciences, University of Technology Sydney

³The ARC Centre of Excellence for Mathematical and Statistical Frontiers

⁴The Alan Turing Institute for Data Science

⁵Department of Engineering Science, University of Oxford

⁶Department of Statistics. University of Oxford

October 21, 2016

Abstract

A research frontier has emerged in scientific computation, wherein numerical error is regarded as a source of epistemic uncertainty that can be modelled. This raises several statistical challenges, including the design of statistical methods that enable the coherent propagation of probabilities through a (possibly deterministic) pipeline of computation. This paper examines the use case for probabilistic numerical methods in routine statistical computation. Our focus is on numerical integration, where a *probabilistic integrator* is equipped with a full distribution over its output that reflects the presence of an unknown numerical error. Our main technical contribution is to establish, for the first time, rates of posterior contraction for these methods. These show that probabilistic integrators can in principle enjoy the “best of both worlds”, leveraging the sampling efficiency of Monte Carlo methods whilst providing a principled route to assessment of the impact of numerical error on scientific conclusions. Several substantial applications are provided for illustration and critical evaluation, including examples from statistical modelling, computer graphics and a computer model for an oil reservoir.

1 Introduction

This paper presents a statistical perspective on the theoretical and methodological issues pertinent to the emerging area of Probabilistic Numerics. Our aim is to stimulate what we feel is an important discussion about these methods for use in contemporary and emerging statistical applications.

Keywords: computational statistics, nonparametric statistics, probabilistic numerics, uncertainty quantification. The authors acknowledge extensive feedback received on an earlier version of this paper from A. Barp, J. Cockayne, J. Dick, D. Duvenaud, A. Gelman, P. Hennig, M. Kanagawa, J. Kronander, X-L. Meng, A. Owen, C. Robert, S. Särkkä, C. Schwab, D. Simpson, J. Skilling, T. Sullivan, Z. Tan, A. Teckentrup and H. Zhu. The authors are very grateful to S. Lan and R. Marques for providing code used in case study 2 and 4. FXB was supported by the EPSRC grant [EP/L016710/1]. CJO was supported by the ARC Centre of Excellence for Mathematical and Statistical Frontiers. MG was supported by the EPSRC grant [EP/J016934/1, EP/K034154/1], an EPSRC Established Career Fellowship, the EU grant [EU/259348] and a Royal Society Wolfson Research Merit Award.

Background Numerical methods, such as linear solvers, quadrature for integration, numerical optimisation methods and discretisation schemes to approximate the solution of differential equations, are the core computational building blocks in modern statistical inference procedures. These are typically considered as computational black-boxes that return a point estimate for a deterministic quantity of interest whose numerical error is then neglected. Numerical methods are thus the only part of the statistical analysis for which uncertainty is not routinely accounted (although analysis of errors and bounds on these are often available and highly developed). In many situations numerical error will be negligible and no further action is required. However, if numerical errors are propagated through a computational pipeline and allowed to accumulate, then failure to properly account for such errors could potentially have drastic consequences on subsequent statistical inferences (Mosbach and Turner, 2009; Conrad et al., 2015).

The study of numerical algorithms from a statistical point of view, where uncertainty is formally due to the presence of an unknown numerical error is known as Probabilistic Numerics. The philosophical foundations for Probabilistic Numerics were, to the best of our knowledge, first clearly exposed in the work of Kadane (1985), Diaconis (1988) and O’Hagan (1992). Theoretical support comes from the foundational work of (Traub et al., 1988; Ritter, 2000), where continuous mathematical operations are approximated by discrete and finite operations to achieve a prescribed accuracy level. Proponents claim that this approach provides three important benefits. Firstly, it provides a principled approach to quantify and propagate numerical uncertainty through computation, allowing for the possibility of errors with complex statistical structure. Secondly, it enables the user to uncover key contributors to numerical error, using statistical techniques such as analysis of variance, then target computational resources at these components. Thirdly, this dual perspective on numerical analysis as an inference task enables new insights, as well as the potential to criticise and refine existing numerical methods. On this final point, recent interest has led to several new and effective numerical algorithms in many areas, including differential equations (Skilling, 1991; Schober et al., 2014; Owhadi, 2016; Cockayne et al., 2016), linear algebra (Hennig, 2015) and optimisation (Hennig and Kiefel, 2013; Mahsereci and Hennig, 2015; Shahriari et al., 2015). These are described in more detail in the recent expositions by Hennig et al. (2015) and the up-to-date literature review provided at www.probabilistic-numerics.org.

Purpose Our aim is to present concrete results that can serve as a basis for discussion on the suitability of probabilistic numerical computation for use in statistical applications. A decision was made to focus on numerical integration (Davis and Rabinowitz, 2007), due to its central role in computational statistics, including frequentist approaches such as bootstrap estimators (Efron and Tibshirani, 1994) and Bayesian approaches, such as computing marginal distributions (Robert and Casella, 2013). In particular we focus on numerical integrals arising in statistics such that the cost of evaluating the integrand forms a computational bottleneck. Such problems occur in statistical analysis for scientific applications involving sophisticated computer simulation, including astrophysics (Sozzetti et al., 2013), meteorology (Mizielinski et al., 2014) and, more broadly, in computing sensitivity estimates, such as Sobol indices (Sobol, 1993) and in uncertainty quantification for inverse problems (Dashti and Stuart, 2016). A probabilistic approach to integration could, if the probabilistic arguments hold, confer several advantages in this context. In particular, from a statistical perspective, valid inferences could be drawn at lower computational budgets by explicitly accounting, in a principled way, for the effect of numerical error on the analysis. Given these substantial claims and a surge in recent research activity on probabilistic numerical methods,

there is now an imperative for the statistical community to critically examine the foundations of uncertainty in deterministic computation (Hennig et al., 2015).

Novel Contribution Consider a probability measure Π on a state space \mathcal{X} . Our task is to compute (or, rather, to *estimate*) integrals of the form

$$\Pi[f] := \int_{\mathcal{X}} f \, d\Pi,$$

where $f : \mathcal{X} \rightarrow \mathbb{R}$ is a test function of interest. Our motivation comes from settings where f does not possess a convenient closed-form expression, so that there is epistemic uncertainty over the actual values attained by f at a point \mathbf{x} until the function is actually evaluated at \mathbf{x} , usually at a non-trivial computational cost.

For numerical integration, the use of a probabilistic numerics approach was first advocated in Kadane and Wasilkowski (1985); Kadane (1985). A few years later, Diaconis (1988) and O’Hagan (1991) developed similar ideas independently. In particular, these previous studies all approach integration from the Bayesian perspective. Whilst not strictly necessary (see Kong et al. (2003) for a maximum likelihood approach), the Bayesian perspective provides an intuitive route to encode information available about the problem at hand.

The probabilistic integration method that we focus on is known as *Bayesian Quadrature* (BQ), the name being coined by O’Hagan (1991). The method operates by evaluating the integrand at a set of states $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$ and returning a probability distribution \mathbb{P}_n , defined over \mathbb{R} , that expresses belief about the true value of $\Pi[f]$. As the name suggests, \mathbb{P}_n will be based on a prior that captures certain properties of f , and that is updated, via Bayes’ rule, on the basis of the “data” contained in the integrand evaluations. The *maximum a posteriori* (MAP) value of \mathbb{P}_n acts as a point estimate of the integral, while the rest of the distribution captures numerical uncertainty due to the fact that we can only evaluate the integrand at a finite number of locations.

Our first contribution is to explore the claim that, if the prior is well-specified (a non-trivial condition), the probability distribution \mathbb{P}_n provides a coherent measure of the uncertainty due to the numerical error. This claim is shown to be substantiated by rigorous mathematical analysis of BQ. In particular the contraction of the posterior \mathbb{P}_n to a point mass centred on the true value $\Pi[f]$ is then established, building on recent work in this direction by Briol et al. (2015). These results hold in the important settings of Markov chain Monte Carlo (MCMC) and Quasi-Monte Carlo (QMC) methods, that are the workhorses of much modern statistical computation.

Our second contribution is to explore the claim that (appropriate) use of prior information can lead to improved rates of convergence, in the usual (non-probabilistic) sense of the point estimate provided by the MAP estimate. Again, we show how this claim can be substantiated with formal analysis of the BQ estimator, in this case provided in the Kernel Quadrature literature (e.g. Sommariva and Vianello, 2006). Going further, we demonstrate that the same conclusion holds for the full BQ posterior; rates of contraction can be improved via the (appropriate) inclusion of prior information. These results serve to highlight the inefficiency of some basic approaches to integration that are often employed in the statistical literature.

Finally, we investigate the extent to which probabilistic integrators are applicable in contemporary statistical applications. In doing so, we have developed strategies for (i) model evidence evaluation via thermodynamic integration, where a large number of candidate models are to be compared, (ii) inverse problems arising in partial differential equation models for oil field fluid flow,

(iii) logistic regression models involving high-dimensional latent random effects, and (iv) spherical integration, as used in the rendering of virtual objects in prescribed environments. In each case the relative advantages and disadvantages of the probabilistic approach to integration are presented for critical evaluation.

Outline The paper is structured as follows. Sec. 2 provides background on BQ and outlines an analytic framework for the method that is rooted in Hilbert spaces. Sec. 3 describes our extension of BQ to MCMC and QMC methods and provides an associated theoretical analysis of their properties. Sec. 4 is devoted to a discussion of practical issues, including the important issue of prior elicitation. Sec. 5 presents several novel applications of probabilistic integration in modern computational statistics. Sec. 6 concludes with a critical appraisal of the suitability of probabilistic numerical methods for statistical computation. Computer codes and scripts in R to reproduce experiments reported in this paper can be downloaded from www.warwick.ac.uk/fxbriol/probabilistic_integration

2 Background

Below we provide the reader with relevant background material. Sec. 2.1 provides a formal description of BQ. Secs. 2.2 and 2.3 explain how the analysis of BQ is dual to minimax analysis in functional regression, and Sec. 2.4 relates these ideas to established MCMC and QMC methods.

Set-Up Let $(\Omega, \mathcal{F}, \mu)$ be a probability space and $(\mathcal{X}, \mathcal{B})$ be a measurable space, whose state space \mathcal{X} will either be a compact subspace of \mathbb{R}^d or more general compact manifolds (e.g. the sphere \mathbb{S}^d), in each case equipped with the Borel σ -algebra $\mathcal{B} = \mathcal{B}(\mathcal{X})$. Also let $\mathbf{X} : \Omega \rightarrow \mathcal{X}$ be a measurable function (random variable). We denote the distribution of \mathbf{X} by Π and consider integration over the measure space $(\mathcal{X}, \mathcal{B}, \Pi)$. Our integrand is assumed to be a measurable function $f : \mathcal{X} \rightarrow \mathbb{R}$ whose expectation, $\Pi[f]$, is the goal of computation.

Notation Write $\|f\|_2^2 := \int_{\mathcal{X}} f^2 d\Pi$ and write $L^2(\Pi)$ for the set of (measurable) functions which are square-integrable with respect to Π (i.e. $\|f\|_2 < \infty$). Note the un-conventional use of $L^2(\Pi)$; here we are *not* referring to equivalence classes of functions. For vector arguments, we also denote $\|\mathbf{u}\|_2 = (u_1^2 + \dots + u_d^2)^{1/2}$. We will make use of the notation $[u]_+ = \max\{0, u\}$. For vector-valued functions $\mathbf{f} : \mathcal{X} \rightarrow \mathbb{R}^m$ or \mathbb{C}^m we write $\Pi[\mathbf{f}]$ for the $m \times 1$ vector whose i th element is $\Pi[f_i]$. The relation $a_l \asymp b_l$ is taken to mean that there exist $0 < C_1, C_2 < \infty$ such that $C_1 a_l \leq b_l \leq C_2 a_l$.

A *quadrature rule* describes any functional $\hat{\Pi} : L^2(\Pi) \rightarrow \mathbb{R}$ that can be written in the linear form

$$\hat{\Pi}[f] = \sum_{i=1}^n w_i f(\mathbf{x}_i), \quad (1)$$

for some states $\{\mathbf{x}_i\}_{i=1}^n \subset \mathcal{X}$ and weights $\{w_i\}_{i=1}^n \subset \mathbb{R}$. The term *cubature rule* is sometimes used when the domain of integration is multi-dimensional (i.e. $d > 1$). The notation $\hat{\Pi}[f]$ is motivated by the fact that this expression can be re-written as the integral of f with respect to an empirical measure $\hat{\Pi} = \sum_{i=1}^n w_i \delta(\cdot - \mathbf{x}_i)$, where $\delta(\cdot)$ is the Dirac delta measure and the weights w_i can be negative and need not have unit sum.

2.1 Bayesian Quadrature

Probabilistic integration begins by defining a prior probability measure over $L^2(\Pi)$. For BQ (O’Hagan, 1991) this is achieved via a Gaussian process (GP) prior (Stein, 1990), motivated in part by analytic tractability but also by the absolute continuity of the posterior measure with respect to the prior measure. A GP is a stochastic process f whose finite dimensional distributions are Gaussian, and can be characterised by its mean function $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$, and its covariance function $k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})])(f(\mathbf{x}') - \mathbb{E}[f(\mathbf{x}')])]$. The common shorthand $\mathcal{GP}(m, k)$ will be employed for these distributions. Note that other prior measures could be taken as the basis for probabilistic integration and may sometimes be more appropriate (e.g. a Student-t process affords heavier tails for values assumed by the integrand).

The next step is to condition the prior measure on (noiseless) “data” $\{(\mathbf{x}_i, f_i)\}_{i=1}^n$, where $f_i = f(\mathbf{x}_i)$, to obtain a posterior measure over $L^2(\Pi)$. For BQ this is also a GP, with mean and covariance denoted m_n and k_n (see e.g. Chap. 2 of Rasmussen and Williams, 2006). The final step is to produce a distribution \mathbb{P}_n on \mathbb{R} by projecting the posterior defined on $L^2(\Pi)$ through the integration operator. A sketch of the procedure is provided in Figure 1 and the relevant formulae are derived below. Denote by $\mathbb{E}_n, \mathbb{V}_n$ the expectation and variance taken with respect to the posterior distribution \mathbb{P}_n . (Later, for measurable A we also write $\mathbb{P}_n[A] = \mathbb{E}_n[1_A]$ where 1_A is the indicator function of the event A .) Write $\mathbf{f} \in \mathbb{R}^n$ for the vector of f_i values, $\mathbf{m} \in \mathbb{R}^n$ for the vector of $m(\mathbf{x}_i)$ values, $X = \{\mathbf{x}_i\}_{i=1}^n$ and $\mathbf{k}(\mathbf{x}, X) = \mathbf{k}(X, \mathbf{x})^T$ for the $1 \times n$ vector whose i th entry is $k(\mathbf{x}, \mathbf{x}_i)$ and \mathbf{K} for the matrix whose i th row and j th column element is $k(\mathbf{x}_i, \mathbf{x}_j)$.

Proposition 1. *The distribution \mathbb{P}_n is Gaussian with*

$$\mathbb{E}_n[\Pi[f]] = \Pi[m] + \Pi[\mathbf{k}(\cdot, X)]\mathbf{K}^{-1}(\mathbf{f} - \mathbf{m}) \quad (2)$$

$$\mathbb{V}_n[\Pi[f]] = \Pi\Pi[k(\cdot, \cdot)] - \Pi[\mathbf{k}(\cdot, X)]\mathbf{K}^{-1}\Pi[\mathbf{k}(X, \cdot)]. \quad (3)$$

Here, $\Pi\Pi[k(\cdot, \cdot)]$ denotes the integral of k with respect to each argument. All proofs in this paper are reserved for Supplement A. Inversion of the matrix \mathbf{K} comes with a cost of $O(n^3)$.

BQ provides a probabilistic model for the epistemic uncertainty over the value of the integral that is due to employing a quadrature rule with a finite number n of function evaluations. By reparametrising $f \mapsto f - m$ we can, without loss of generality, suppose that $m \equiv 0$ for the remainder of the paper. Then, from Eqn. 2, the mean takes the form of a quadrature rule

$$\mathbb{E}_n[\Pi[f]] = \hat{\Pi}_{\text{BQ}}[f] := \sum_{i=1}^n w_i^{\text{BQ}} f(\mathbf{x}_i) \quad (4)$$

where $\mathbf{w}^{\text{BQ}} := \mathbf{K}^{-1}\Pi[\mathbf{k}(X, \cdot)]$. Furthermore, from Eq. 3 the variance $\mathbb{V}_n[\Pi[f]]$ does not depend on function values $\{f_i\}_{i=1}^n$, but only on the location of the states $\{\mathbf{x}_i\}_{i=1}^n$ and the choice of covariance function k . This is useful as it allows state locations and weights to be pre-computed and re-used to integrate multiple integrals under the same prior specification. However, it also means that the variance is driven by the choice of prior on $L^2(\Pi)$, which will have great influence on the performance of the method. A valid quantification of uncertainty relies on a well-specified prior, and we will consider this issue further in Sec. 4.1.

It has long been known that the BQ mean (Eqn. 4) coincides with classical quadrature rules for specific choices of covariance function k . For example, a Brownian covariance function k leads to a posterior mean m_n that is a piecewise linear interpolation of f between the states $\{\mathbf{x}_i\}_{i=1}^n$,

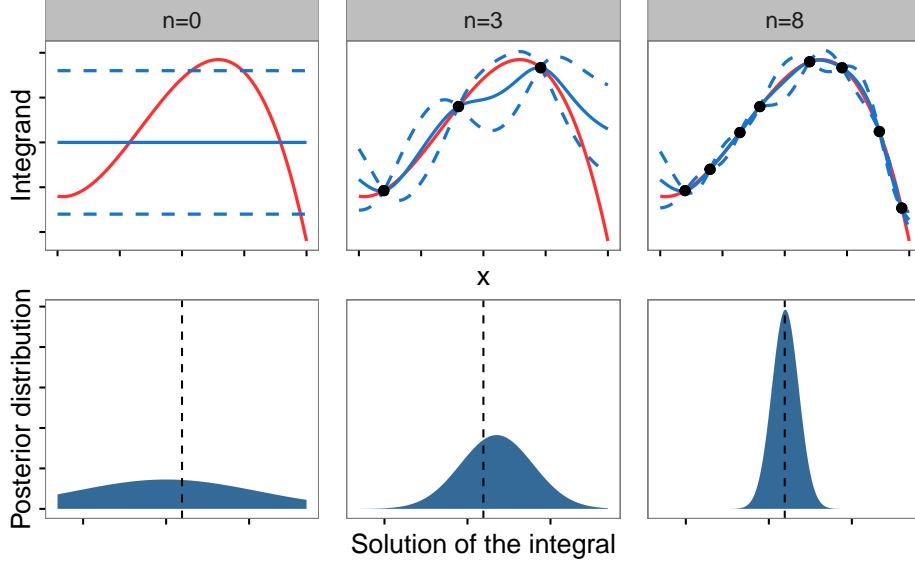


Figure 1: Sketch of Bayesian Quadrature. The top row shows the approximation of the integrand f (in red) by the posterior mean m_n (in blue) as the number n of function evaluations is increased. The dashed lines represent 95% posterior credible intervals. The bottom row shows the Gaussian distribution with mean $\mathbb{E}_n[\Pi[f]]$ and variance $\mathbb{V}_n[\Pi[f]]$ and the dashed black line gives the true value of the integral $\Pi[f]$. As the number of states n increased, the the posterior distribution \mathbb{P}_n over \mathbb{R} contracts onto the true value of the integral $\Pi[f]$.

i.e. the trapezium rule (Suldin, 1959). As another example, an integrated Brownian covariance function k results in a cubic spline interpolant for m_n , i.e. an instance of Simpson’s rule (Diaconis, 1988). Clearly the point estimator in Eqn. 4 is a natural object; it has also received attention in both the Kernel Quadrature literature (Sommariva and Vianello, 2006) and Empirical Interpolation literature (Kristoffersen, 2013). From the probabilistic perspective, after O’Hagan (1991), subsequent contributions to the BQ literature focus on computational considerations and include Minka (2000); Rasmussen and Ghahramani (2002); Huszar and Duvenaud (2012); Gunter et al. (2014) and Briol et al. (2015). The present paper focuses both on theoretical and empirical results and provides substantial novel contributions in each respect.

An important motivation for BQ is the case of expensive integrands f , where the cost of matrix inversion is provided in Supplement C.

2.2 Quadrature Rules in Hilbert Spaces

Below we describe how analysis of the approximation properties of the quadrature rule $\hat{\Pi}_{\text{BQ}}[f]$ can be carried out in terms of function spaces, and in particular in terms of reproducing kernel Hilbert space (RKHS; Berlinet and Thomas-Agnan, 2004). This can be considered as necessary preparation for analysis of the full distribution \mathbb{P}_n in Sec. 3.

Consider a Hilbert space \mathcal{H} with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and associated norm $\| \cdot \|_{\mathcal{H}}$. \mathcal{H} is said to be an RKHS if there exists a symmetric, positive definite function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, called a *kernel*, that satisfies two properties: (i) $k(\cdot, \mathbf{x}) \in \mathcal{H}$ for all $\mathbf{x} \in \mathcal{X}$ and; (ii) $f(\mathbf{x}) = \langle f, k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}}$ for

all $\mathbf{x} \in \mathcal{X}$ and $f \in \mathcal{H}$ (the *reproducing* property). It can be shown that every kernel defines an RKHS and every RKHS admits a unique reproducing kernel (Berlinet and Thomas-Agnan, 2004, Sec. 1.3). In this paper all kernels k are assumed to satisfy $R := \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}) \Pi(d\mathbf{x}) < \infty$. In particular this guarantees $f \in L^2(\Pi)$ for all $f \in \mathcal{H}$. Define the *kernel mean* map $\mu(\Pi) : \mathcal{X} \rightarrow \mathbb{R}$ as $\mu(\Pi)(\mathbf{x}) := \Pi[k(\cdot, \mathbf{x})]$. This exists in \mathcal{H} as an implication of the assumption $R < \infty$ (Smola et al., 2007). The name is justified by the fact that, for all $f \in \mathcal{H}$:

$$\begin{aligned} \Pi[f] &= \int_{\mathcal{X}} f \, d\Pi = \int_{\mathcal{X}} \langle f, k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}} \Pi(d\mathbf{x}) \\ &= \left\langle f, \int_{\mathcal{X}} k(\cdot, \mathbf{x}) \Pi(d\mathbf{x}) \right\rangle_{\mathcal{H}} = \langle f, \mu(\Pi) \rangle_{\mathcal{H}}. \end{aligned}$$

Here the integral and inner product commute due to the existence of $\mu(\Pi)$ as a Bochner integral over an RKHS (Steinwart and Christmann, 2008, p510). The reproducing property permits an elegant theoretical analysis of quadrature rules, with many quantities of interest tractable analytically in \mathcal{H} . In the language of kernel means, quadrature rules of the form in Eqn. 1 can be written in the form $\hat{\Pi}[f] = \langle f, \mu(\hat{\Pi}) \rangle_{\mathcal{H}}$ where $\mu(\hat{\Pi})$ is the approximation to the kernel mean given by $\mu(\hat{\Pi})(\mathbf{x}) = \hat{\Pi}[k(\cdot, \mathbf{x})]$. For fixed $f \in \mathcal{H}$, the integration error associated with $\hat{\Pi}$ can then be expressed as

$$\hat{\Pi}[f] - \Pi[f] = \langle f, \mu(\hat{\Pi}) \rangle_{\mathcal{H}} - \langle f, \mu(\Pi) \rangle_{\mathcal{H}} = \langle f, \mu(\hat{\Pi}) - \mu(\Pi) \rangle_{\mathcal{H}}.$$

An upper bound for the error is obtained by applying the Cauchy-Schwarz inequality:

$$|\hat{\Pi}[f] - \Pi[f]| \leq \|f\|_{\mathcal{H}} \|\mu(\hat{\Pi}) - \mu(\Pi)\|_{\mathcal{H}}. \quad (5)$$

The expression above decouples the magnitude (in \mathcal{H}) of the integrand f from the approximation accuracy of the kernel mean. The following sections discuss how quadrature rules can be tailored to target the second term in this upper bound on the integration error.

2.3 Optimality of Bayesian Quadrature Weights

Denote the dual space of \mathcal{H} , consisting of all bounded linear functionals $\mathcal{H} \rightarrow \mathbb{R}$, by \mathcal{H}^* and denote its corresponding norm $\|\cdot\|_{\mathcal{H}^*}$. The performance of quadrature rules can be quantified by the *worst-case error* (WCE) in the RKHS:

$$\|\hat{\Pi} - \Pi\|_{\mathcal{H}^*} = \sup_{\|f\|_{\mathcal{H}} \leq 1} |\hat{\Pi}[f] - \Pi[f]|$$

The WCE is characterised as the error in estimating the kernel mean:

Proposition 2. $\|\hat{\Pi} - \Pi\|_{\mathcal{H}^*} = \|\mu(\hat{\Pi}) - \mu(\Pi)\|_{\mathcal{H}}$.

Minimisation of the WCE is natural and corresponds to solving a least-squares problem in the feature space induced by the kernel; such a solution gives minimax properties in the original space (Ritter, 2000, Prop. III.17). This least-squares formulation is analytically tractable: Letting $\mathbf{w} \in \mathbb{R}^n$ denote the vector of weights $\{w_i\}_{i=1}^n$, $\mathbf{z} \in \mathbb{R}^n$ be a vector such that $z_i = \mu(\Pi)(\mathbf{x}_i)$, and $\mathbf{K} \in \mathbb{R}^{n \times n}$ be the matrix with entries $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$, we obtain the following from direct calculation:

Proposition 3. $\|\hat{\Pi} - \Pi\|_{\mathcal{H}^*}^2 = \mathbf{w}^T \mathbf{K} \mathbf{w} - 2\mathbf{w}^T \mathbf{z} + \Pi[\mu(\Pi)]$.

Several optimality properties for integration in RKHS were collated in Sec. 4.2 of Novak and Woźniakowski (2008). Relevant to this work is that an optimal (i.e. minimax) estimate $\hat{\Pi}$ can, without loss of generality, take the form of a quadrature rule (i.e. of the form $\hat{\Pi}$ in Eqn. 1). To be more precise, any non-linear estimator *adaptive* estimator, which infers f “on-the-fly”, can be matched in terms of WCE by a (linear) quadrature rule as defined above. (Of course, adaptive quadratures may provide superior performance for a *single* fixed function f , and the minimax result is not true in general outside the RKHS framework.)

To relate these ideas to BQ, consider the challenge of deriving an optimal quadrature rule, conditional on fixed states $\{\mathbf{x}_i\}_{i=1}^n$, that minimises the WCE over weights $\mathbf{w} \in \mathbb{R}^n$. From Prop. 3, the solution to this convex problem is easily seen to be $\mathbf{w} = \mathbf{K}^{-1}\mathbf{z}$. This shows that the BQ point estimate, based on a prior covariance function k , is identical to the *optimal* quadrature rule in the RKHS whose reproducing kernel is k (Kadane and Wasilkowski, 1985; Ritter, 2000). (Of course, this is a particular instance of the more general result that Bayes estimates are minimax.) Furthermore, the expression for the WCE in Prop. 3 shows that, for any other quadrature rule $\hat{\Pi}$ based on the same states $\{\mathbf{x}_i\}_{i=1}^n$,

$$\mathbb{V}_n[\Pi[f]] = \|\hat{\Pi}_{\text{BQ}} - \Pi\|_{\mathcal{H}^*}^2 \leq \|\hat{\Pi} - \Pi\|_{\mathcal{H}^*}^2$$

with equality if and only if the BQ weights are employed for $\hat{\Pi}$. Regarding optimality, the problem is thus reduced to selection of states $\{\mathbf{x}_i\}_{i=1}^n$.

Our analysis in this paper is reliant on the (strong) assumption that the integrand f belongs to the RKHS \mathcal{H} ; in particular, this is stronger than the assumption that the integrand f is in the support of the prior measure. Further discussion is provided in Sec. 6.

2.4 Selection of States

In earlier work, O’Hagan (1991) used the same states $\{\mathbf{x}_i\}_{i=1}^n$ that are employed in Gaussian quadrature formulae. Later Rasmussen and Ghahramani (2002) generated states using Monte Carlo (MC), calling the approach *Bayesian* MC (BMC). Recent work by Gunter et al. (2014); Briol et al. (2015) selected states using experimental design by targeting the variance $\mathbb{V}_n[\Pi[f]]$. These approaches are briefly recalled below.

2.4.1 Monte Carlo Methods

An MC method is a quadrature rule based on uniform weights $w_i^{\text{MC}} := 1/n$ and states $\{\mathbf{x}_i\}_{i=1}^n$ that are considered as random variables. The simplest of those methods consists of sampling states $\{\mathbf{x}_i^{\text{MC}}\}_{i=1}^n$ independently from Π . For un-normalised densities, MCMC methods proceed similarly but induce a dependence structure among the $\{\mathbf{x}_i^{\text{MCMC}}\}_{i=1}^n$. We denote these (random) estimators by $\hat{\Pi}_{\text{MC}}$ (when $\mathbf{x}_i = \mathbf{x}_i^{\text{MC}}$) and $\hat{\Pi}_{\text{MCMC}}$ (when $\mathbf{x}_i = \mathbf{x}_i^{\text{MCMC}}$) respectively. Uniformly weighted estimators are well-suited to many challenging integration problems since they provide a dimension-independent convergence rate for the WCE of $O_P(n^{-1/2})$ (Thm. 1 below). They are also widely applicable and straight-forward to analyse; for instance the central limit theorem (CLT) gives that $\sqrt{n}(\hat{\Pi}_{\text{MC}}[f] - \Pi[f]) \rightarrow \mathcal{N}(0, \tau_f^{-1})$ where $\tau_f^{-1} = \Pi[f^2] - \Pi[f]^2$ and the convergence is in distribution. However, the CLT is not well-suited as a measure of *epistemic* uncertainty (i.e. as an explicit model for numerical error) since (i) it is only valid asymptotically, and (ii) τ_f is unknown, depending on the integral $\Pi[f]$ that we are attempting to estimate.

QMC methods exploit knowledge of the RKHS \mathcal{H} to spread the states in an efficient, deterministic way over the domain \mathcal{X} . QMC also approximates integrals using a quadrature rule $\hat{\Pi}_{\text{QMC}}[f]$ that has uniform weights $w_i^{\text{QMC}} := 1/n$. These methods benefit from an extensive theoretical literature (Hickernell, 1998). The (in some cases) optimal convergence rates, as well as sound statistical properties, of QMC have recently led to interest within statistics (e.g. Hickernell et al., 2005; Gerber and Chopin, 2015; Oates et al., 2016c).

2.4.2 Experimental Design Methods

An *Optimal* BQ (OBQ) rule selects states $\{\mathbf{x}_i\}_{i=1}^n$ to globally minimise the variance $\mathbb{V}_n[\Pi[f]]$. OBQ corresponds to classical quadrature rules (e.g. Gauss-Hermite) for specific choices of kernels (Särkka et al., 2016). However OBQ can generally not be implemented; the problem of optimizing states is in general NP-hard (Schölkopf and Smola, 2002, Sec. 10.2.3).

A more pragmatic approach to select states is using experimental design methods, such as the greedy algorithm sequentially minimising $\mathbb{V}_n[\Pi[f]]$ at each iteration. This method, called *Sequential* BQ (SBQ), is straightforward to implement, e.g. using general-purpose numerical optimisation, and is a probabilistic integration method that is often used in practice (Osborne et al., 2012; Gunter et al., 2014). More sophisticated optimisation algorithms have also been used to select states. For example, in the QMC literature Nuyens and Cools (2006) framed the construction of lattice rules as an optimisation problem, in the least-squares approximation literature Maday et al. (2009) proposed a greedy algorithm to generate approximate Fekete states. In the Empirical Interpolation literature Eftang and Stamm (2012) proposed adaptive procedures that iteratively divide the domain of integration into sub-domains. In the BQ literature Briol et al. (2015) used conditional gradient algorithms (also called Frank-Wolfe algorithms) to select states. Finally, we point out that Monte Carlo methods based on determinantal point processes (Bardenet and Hardy, 2016) could also provide an efficient alternative to the design problem.

At present, experimental design schemes do not possess the computational efficiency that we have come to expect from advanced MCMC and QMC methods. Moreover, they do not scale well to high-dimensional settings due to the need to repeatedly solve high-dimensional optimisation problems and have few established theoretical guarantees. For these reasons we focused on MC, MCMC and QMC below.

3 Methods

This section presents novel theoretical results on probabilistic integration methods in which the states $\{\mathbf{x}_i\}_{i=1}^n$ are the output of MCMC and QMC algorithms. Sec. 3.1 provides formal definitions, while Sec. 3.2 establishes theoretical properties.

3.1 Probabilistic Integration

The sampling methods of MCMC (Robert and Casella, 2013) and, to a lesser extent, QMC (Dick and Pillichshammer, 2010) are central to much of contemporary statistical computation. Here we pursue the idea of using these methods to generate states for BQ, with the aim to exploit BQ within statistical applications to account for the possible impact of numerical error on inferences. More precisely, this paper studies the two-step procedure that first uses MCMC or QMC in order to select states and then assigns BQ weights to those states. In the MCMC case it is possible that

two states $\mathbf{x}_i = \mathbf{x}_j$ are identical, for example in a Metropolis-Hastings chain. To prevent the kernel matrix \mathbf{K} from becoming singular, duplicate states should be discarded. (This is justified since the information contained in function evaluations $f_i = f_j$ is not lost. This does *not* introduce bias into BQ methods, in contrast to MC methods.) Thus we define

$$\begin{aligned}\hat{\Pi}_{\text{BMCMC}}[f] &:= \sum_{i=1}^n w_i^{\text{BQ}} f(\mathbf{x}_i^{\text{MCMC}}) \\ \hat{\Pi}_{\text{BQMC}}[f] &:= \sum_{i=1}^n w_i^{\text{BQ}} f(\mathbf{x}_i^{\text{QMC}})\end{aligned}$$

This two-step procedure requires no modification to existing MCMC or QMC sampling schemes. Each estimator is associated with a full posterior probability distribution \mathbb{P}_n described in Sec. 2.1.

A moment is taken to emphasise that the apparently simple act of re-weighting MCMC samples can have a dramatic improvement on convergence rates in estimation. This is a well-known fact from the Kernel Quadrature literature and we point the interested reader toward Sommariva and Vianello (2006). Whilst our main interest is in the suitability of BQ as a statistical model for numerical integration error, we first also highlight the efficient point estimation which comes out as a by-product of the probabilistic numerics approach.

To date we are not aware of any previous use of BMCMC, presumably due to analytic intractability of the kernel mean; we provide one possible solution in the following sections. BQMC has been described by Hickernell et al. (2005); Marques et al. (2013); Särkka et al. (2016). To the best of our knowledge there has been no theoretical analysis of the distributions \mathbb{P}_n associated with either method. The goal of the next section is to establish theoretical guarantees for consistency of these point estimators and contraction of their associated posterior distributions.

3.2 Theoretical Properties

We have now completed our review of probabilistic numerics approaches to integration. In this section we present novel theoretical results for BMC, BMCMC and BQMC. Our primary focus here is on integration in Sobolev spaces. This is due to the fact that numerical conditioning is known to be better behaved in these spaces compared to spaces of smoother functions (Schaback, 1995). In addition, there is an elegant characterisation of these spaces in terms of the number of (weak) derivatives of the integrand. Thus from a practical perspective, it is straight-forward, at least in principle, to select a suitable space of functions in which to work.

3.2.1 Bayesian Markov Chain Monte Carlo

As a baseline, we begin by noting a general result for MC estimation. This requires a slight strengthening of the assumption on the kernel: $k_{\max} := \sup_{\mathbf{x} \in \mathcal{X}} k(\mathbf{x}, \mathbf{x})^{1/2} < \infty$. This implies that f is bounded on \mathcal{X} . For MC estimators, Lemma 33 of Song (2008) show that, when $k_{\max} < \infty$, the WCE converges in probability at the classical root- n rate:

Lemma 1 (MC Methods). $\|\hat{\Pi}_{\text{MC}} - \Pi\|_{\mathcal{H}^*} = O_P(n^{-1/2})$.

MC estimates are widely used in statistical applications, in part due to its weak assumption of $f \in L^2(\Pi)$.

Turning now to BMCMC (and BMC as a special case), we obtain rates for the WCE that improve on the MC rate in certain cases. Consider the compact manifold $\mathcal{X} = [0, 1]^d$. Below the measure Π will be assumed to admit a density with respect to Lebesgue measure, denoted by $\pi : \mathcal{X} \rightarrow [0, \infty)$. Write \mathfrak{F} for the Fourier transform operator and define the *Sobolev* space $\mathcal{H}_\alpha := \{f \in L^2(\Pi) \text{ such that } \|f\|_{\mathcal{H}, \alpha} < \infty\}$, equipped with the norm $\|f\|_{\mathcal{H}, \alpha} := \|\mathfrak{F}^{-1}[(1 + \|\boldsymbol{\xi}\|_2^2)^{\alpha/2} \mathfrak{F}[f]]\|_2$. Here α is the *order* of the space. It can be shown that \mathcal{H}_α is the set of functions f whose weak derivatives $(\partial x_1)^{u_1} \dots (\partial x_d)^{u_d} f$ exist in $L^2(\Pi)$ for $u_1 + \dots + u_d \leq \alpha$ (Triebel, 1992).

To emphasise the level of generality here: a function with derivatives up to order α belongs to \mathcal{H}_α . Derivative counting can hence be a principled approach for practitioners to choose a kernel. In fact, our results are more general than Sobolev spaces, in the sense that any radial kernel $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x} - \mathbf{x}')$ whose Fourier transform decays at a rate $\mathfrak{F}[\phi(\boldsymbol{\xi})] \asymp (1 + \|\boldsymbol{\xi}\|_2^2)^{-\alpha}$, where $\alpha > d/2$, (e.g. Matérn kernel) induces an RKHS that is norm-equivalent to the standard Sobolev space \mathcal{H}_α (Wendland, 2005, Cor. 10.13). (Two norms $\|\cdot\|$, $\|\cdot\|'$ on a vector space \mathcal{H} are *equivalent* when there exists constants $0 < C_1, C_2 < \infty$ such that for all $h \in \mathcal{H}$ we have $C_1 \|h\| \leq \|h\|' \leq C_2 \|h\|$.) All results below apply to RKHS \mathcal{H} that are norm-equivalent to \mathcal{H}_α , permitting flexibility in the choice of kernel. Specific examples of kernels are provided in Sec. 4.2.

Our analysis below is based on the scattered data approximation literature (Wendland, 2005). A minor technical assumption, that enables us to simplify the presentation of results below, is that the set $X = \{\mathbf{x}_i\}_{i=1}^n$ may be augmented with a finite, pre-determined set $Y = \{\mathbf{y}_i\}_{i=1}^m$ where m does not increase with n . Clearly this has no bearing on asymptotics. Define $I_D = [\Pi[f] - D, \Pi[f] + D]$ to be an interval of radius D centred on the true value of the integral. We are interested in $\mathbb{P}_n[I_D^c]$, the posterior mass on $I_D^c = \mathbb{R} \setminus I_D$, and in particular how quickly this mass vanishes as the number n of function evaluations is increased.

Theorem 1 (BMCMC in \mathcal{H}_α). *Suppose π is bounded away from zero on \mathcal{X} . Let \mathcal{H} be norm-equivalent to \mathcal{H}_α where $\alpha > d/2$, $\alpha \in \mathbb{N}$. Suppose states are generated by a reversible, uniformly ergodic Markov chain that targets Π . Then*

$$\|\hat{\Pi}_{BMCMC} - \Pi\|_{\mathcal{H}^*} = O_P(n^{-\alpha/d+\epsilon})$$

and

$$\mathbb{P}_n[I_D^c] = O_P(\exp(-Cn^{2\alpha/d-\epsilon})),$$

where $C > 0$ depends on D and $\epsilon > 0$ can be arbitrarily small.

This result shows the posterior distribution is well-behaved; probability mass concentrates on the true solution as n increases. Hence, if our prior is well-specified (see Sec. 4.1), the posterior distribution provides valid uncertainty quantification over the solution of the integral as a result of performing a finite number n of function evaluations. This result does not address the coverage of \mathbb{P}_n , which is assessed empirically in Sec. 5.

An information-theoretic lower bound on the WCE in this setting is $O_P(n^{-\alpha/d-1/2})$ (Novak and Woźniakowski, 2010). Thus our results here are at most one MC rate away from being optimal. Bach (2015) obtained a similar result for fixed n and specific importance sampling distribution (often intractable in practice); his analysis does not directly imply our asymptotic results and *vice versa*.

The results for BMC and BMCMC point estimation demonstrate the improved accuracy relative to MC, supporting its use in the case of expensive integration problems. It would not be practical

in settings where the number n of MCMC samples is large, due to the $O(n^3)$ cost, however in such settings the contribution from numerical error is likely to be negligible. As a default STAN (Carpenter et al., 2016) stores $n = 4000$ samples; this is about the limit for BQ on a modern computer without recourse to more sophisticated linear algebraic methods.

Thm. 1 can be generalised in several directions. Firstly, we can consider more general domains \mathcal{X} . Specifically, the scattered data approximation bounds that are used in our proof apply to any compact domain $\mathcal{X} \subset \mathbb{R}^d$ that satisfies an *interior cone condition* (Wendland, 2005, p.28). In addition, we can consider other spaces \mathcal{H} , such as the RKHS obtained from popular Gaussians, multi-quadrics, inverse multi-quadrics and thin-plate splines kernels. For example, an extension of Thm. 1 shows that the Gaussian kernel leads to exponential rates for the WCE and super-exponential rates for posterior contraction. We refer the interested reader to the extended results available in Supplement B.

3.2.2 Bayesian Quasi Monte Carlo

The previous section showed that BMCMC is close to rate-optimal in \mathcal{H}_α . To avoid repetition, here we consider more interesting spaces of functions whose *mixed* partial derivatives exist, for which even faster convergence rates can be obtained using BQMC methods. To formulate BQMC we must posit an RKHS *a priori* (i.e. specifying all kernel parameters) and consider collections of states $\{\mathbf{x}_i^{\text{QMC}}\}_{i=1}^n$ that constitute a QMC point set tailored to the RKHS.

Consider the compact manifold $\mathcal{X} = [0, 1]^d$ with Π be uniform on \mathcal{X} . Define the Sobolev space of *dominating mixed smoothness* $\mathcal{S}_\alpha := \{f \in L^2(\Pi) \text{ such that } \|f\|_{\mathcal{S}, \alpha} < \infty\}$, equipped with the norm $\|f\|_{\mathcal{S}, \alpha} := \|\mathfrak{F}^{-1}[\prod_{i=1}^d (1 + \xi_i^2)^{\alpha/2} \mathfrak{F}[f]]\|_2$. Here α is the *order* of the space. It can be shown that \mathcal{S}_α is the set of functions f whose weak derivatives $(\partial x_1)^{u_1} \dots (\partial x_d)^{u_d} f$ exist in $L^2(\Pi)$ for $u_i \leq \alpha$, $i = 1, \dots, d$. To build intuition, note that \mathcal{S}_α is norm-equivalent to the RKHS generated by a tensor product of Matérn kernels (Sickel and Ullrich, 2009), or indeed a tensor product of any other univariate Sobolev space -generating kernel.

For a specific space such as \mathcal{S}_α , we seek an appropriate QMC point set. The *higher-order digital* $(t, \alpha, 1, \alpha m \times m, d)$ -*net* construction is an example of a QMC point set for \mathcal{S}_α . Intuitively, digital nets aim to spread states evenly across the domain \mathcal{X} in order to obtain efficient (un-weighted) quadrature rules; we refer the reader to Dick and Pillichshammer (2010) for details of this construction.

Theorem 2 (BQMC in \mathcal{S}_α). *Let \mathcal{H} be norm-equivalent to \mathcal{S}_α , where $\alpha \geq 2$, $\alpha \in \mathbb{N}$. Suppose states are chosen according to a higher-order digital $(t, \alpha, 1, \alpha m \times m, d)$ net over \mathbb{Z}_b for some prime b where $n = b^m$. Then*

$$\|\hat{\Pi}_{\text{BQMC}} - \Pi\|_{\mathcal{H}^*} = O(n^{-\alpha+\epsilon})$$

and

$$\mathbb{P}_n[I_D^c] = O(\exp(-Cn^{2\alpha-\epsilon})),$$

where $C > 0$ depends on D and $\epsilon > 0$ can be arbitrarily small.

This is the *optimal* rate for any deterministic quadrature rule in \mathcal{S}_α (Novak and Woźniakowski, 2010). These results should be understood to hold on the sub-sequence $n = b^m$, as QMC methods do not in general give guarantees for all $n \in \mathbb{N}$. It is not clear how far this result can be generalised, in terms of Π and \mathcal{X} , compared to the result for BMCMC, since this would require the use of different QMC point sets. The case of QMC for the Gaussian kernel was recently studied in Fasshauer et al.

(2012); the results therein for Smolyak point sets imply (exponential) convergence and contraction rates for BQMC in via the same arguments that we have made explicit for the space \mathcal{S}_α in this section.

3.2.3 Summary

This concludes the presentation of our main theoretical results. We have focused on establishing near-optimal rates of convergence and contraction for BMC, BMCMC and BQMC in general Sobolev space settings. These results are essential since they establish the sound properties of the probability measure \mathbb{P}_n , which is shown to contract to the truth as more evaluations are made of the integrand. Moreover, kernel quadrature rules are shown to be asymptotically efficient relative to the simple MC estimator that is widely used in statistical applications. Of course, the higher computational cost of up to $O(n^3)$ may sometime restrict the applicability of the method in large- n regimes. However, we emphasise that the main aim of probabilistic numerics methods is to quantify the uncertainty induced from numerical errors, an important task which often justifies the higher computational cost. Before exploring these estimators on concrete examples, the next section discusses methodological considerations relevant to the implementation of these methods, including specification of the prior distribution.

4 Implementation

So far we have established sound theoretical properties for BMCMC and BQMC under the assumption that the prior is well-specified. Unfortunately, prior specification complicates the situation in practice since, given a test function f , there are an infinity of RKHS that f belongs to and the specific choice of this space will impact upon the performance of the method. In particular, the scale of the prior will have a significant bearing on the suitability of the uncertainty quantification and could lead to over-confident inferences, which mitigates the advantages of the probabilistic numerical framework. Below we discuss this, along with a number of other practical considerations.

4.1 Prior Specification

The theoretical results above deal with asymptotic behaviour, but a question remains over the choice of prior in relation to the performance afforded for finite n . i.e. whether the scale of the posterior uncertainty provides an accurate reflection of the actual numerical error in practice. For BQ, this corresponds to the well-studied problem of prior specification in the kriging literature (Stein, 1990) or of kernel learning in machine learning (Duvenaud, 2014).

At this point we highlight a distinction between B(MC)MC and both BQMC and experimental design BQ schemes; for the former the choice of states does not depend on the RKHS. For B(MC)MC this allows for the possibility of off-line specification of the kernel after evaluations of the integrand have been obtained, whereas for alternative methods the kernel must be stated up-front. Our discussion below therefore centres on prior specification in relation to B(MC)MC, where statistical techniques offer several resolutions.

Consider a parametric kernel $k(\mathbf{x}, \mathbf{x}'; \theta_l, \theta_s)$, with a distinction drawn here between *scale* parameters θ_l and *smoothness* parameters θ_s . The former are defined as parametrising the norm on \mathcal{H} , whereas the latter affect the set \mathcal{H} itself. Calibration based on data can only be successful in

the absence of acute sensitivity to these parameters. For scale parameters a wide body of empirical evidence demonstrates that this is usually not a concern, but this is not true for smoothness parameters. Indeed, selection of smoothness parameters is an active area of theoretical research (e.g. Szabó et al., 2015). In some cases it is possible to elicit a smoothness parameter from physical or mathematical considerations, such as the number of derivatives that the integrand is known to possess. Our attention below is focused on calibrating kernel scale parameters. Several possible approaches are discussed below in relation to their suitability for BQ.

Marginalisation A natural approach, from a Bayesian perspective, is to set a prior $p(\theta_l)$ on parameters θ_l and then to marginalise over θ_l to obtain a posterior over $\Pi[f]$. Recent results for the Gaussian kernel establish minimax optimal adaptive rates in $L^2(\Pi)$ for this approach, including in the practically relevant setting where Π is supported on a low-dimensional sub-manifold of the ambient space \mathcal{X} (Yang and Dunson, 2013). However, the act of marginalisation itself involves an intractable integral. While the computational cost of evaluating this integral will often be dwarfed by that of the integral $\Pi[f]$ of interest, marginalisation nevertheless introduces an additional undesirable computational challenge that might require several approximations (e.g. Osborne, 2010).

Cross-Validation One approach to kernel choice would be k -fold cross-validation. However, this can perform poorly when the number n of data is small, since the data needs to be further reduced into training and test sets. The performance estimates are also known to have large variance in those cases (Chap. 5 of Rasmussen and Williams, 2006). Since the small n scenario is one of our primary settings of interest for BQ, we felt that cross-validation was unsuitable for use in applications below.

Empirical Bayes An alternative to the above approaches is *empirical Bayes* (EB) selection of scale parameters, choosing θ_l to maximise the log-marginal likelihood of the data $f(\mathbf{x}_i)$, $i = 1, \dots, n$ (Sec. 5.4.1 of Rasmussen and Williams, 2006). EB has the advantage of providing an objective function that is easier to optimise relative to cross-validation. However, we also note that EB can lead to over-confidence when n is very small, since the full irregularity of the integrand has yet to be uncovered (Szabó et al., 2015).

For the remainder, we chose to focus on the EB approach. Empirical results support the use of EB here, though we do not claim that this is an optimal strategy and we invite further investigation.

4.2 Tractable and Intractable Kernel Means

BQ requires that the kernel mean $\mu(\Pi)(\mathbf{x}) = \Pi[k(\cdot, \mathbf{x})]$ is available in closed-form. This is the case for several kernel-distribution pairs (k, Π) and a subset of these pairs are recorded in Table 1. These pairs are widely applicable; for example the gradient-based kernel in Oates et al. (2016a,b) provides a closed-form expression for the kernel mean whenever the gradient $\partial \log \pi(\mathbf{x})$ is available, and is compatible with un-normalised densities that arise in MCMC sampling.

In the event that the kernel-distribution pair (k, Π) of interest does not lead to a closed-form kernel mean, it is sometimes possible to determine another kernel-density pair (k', Π') for which $\Pi'[k'(\cdot, \mathbf{x})]$ is available and such that (i) Π is absolutely continuous with respect to Π' , so that the Radon-Nikodym derivative $d\Pi/d\Pi'$ exists, and (ii) $f d\Pi/d\Pi' \in \mathcal{H}(k')$. Then one can construct an

\mathcal{X}	Π	k	Reference
$[0, 1]^d$	Unif(\mathcal{X})	Wendland TP	Oates et al. (2016c)
$[0, 1]^d$	Unif(\mathcal{X})	Matérn Weighted TP	Sec. 5.2.3
$[0, 1]^d$	Unif(\mathcal{X})	Exponentiated Quadratic	Use of error function
\mathbb{R}^d	Mixt. of Gaussians	Exponentiated Quadratic	O’Hagan (1991)
\mathbb{S}^d	Unif(\mathcal{X})	Gegenbauer	Sec. 5.2.4
Arbitrary	Unif(\mathcal{X}) / Mixt. of Gauss.	Trigonometric	Integration by parts
Arbitrary	Unif(\mathcal{X})	Splines	Wahba (1990)
Arbitrary	Known moments	Polynomial TP	Briol et al. (2015)
Arbitrary	Known $\partial \log \pi(\mathbf{x})$	Gradient-based Kernel	Oates et al. (2016a,b)

Table 1: A non-exhaustive list of distribution Π and kernel k pairs that provide a closed-form expression for both the kernel mean $\mu(\Pi)(\mathbf{x}) = \Pi[k(\cdot, \mathbf{x})]$ and the initial error $\Pi[\mu(\Pi)]$. Here TP refers to the tensor product of one-dimensional kernels.

importance sampling estimator

$$\Pi[f] = \int_{\mathcal{X}} f \, d\Pi = \int_{\mathcal{X}} f \frac{d\Pi}{d\Pi'} \, d\Pi' = \Pi' \left[f \frac{d\Pi}{d\Pi'} \right] \quad (6)$$

and proceed as above (O’Hagan, 1991).

One side-contribution of this research was a novel and generic approach to accommodate intractability of the kernel mean in BQ. This is described in detail in Supplement C.4 and used in case studies #1 and #2 presented in Sec. 5. A second, more Bayesian solution was proposed in Oates et al. (2016d).

5 Results

The aims of the following section are two-fold; (i) to validate the preceding theoretical analysis and (ii) to explore the use of probabilistic integrators in a range of problems arising in contemporary statistical applications.

5.1 Assessment of Uncertainty Quantification

Our focus below is on the uncertainty quantification provided by BQ and, in particular, the performance of the EB prior specification. To be clear, we are not concerned with accurate point estimation at low computational cost. This is a well-studied problem that reaches far beyond the methods of this paper (e.g. Deylon and Portier, 2016). Rather, we are aiming to assess the suitability of the probabilistic description for numerical error that is provided by these probabilistic integrators. Our motivation is expensive integrands, but to perform assessment in a controlled environment we considered inexpensive test functions of varying degrees of irregularity, whose integrals can be accurately approximated by brute force. These included an “easy” test function $f_1(\mathbf{x}) = \exp(\sin(5\|\mathbf{x}\|_2)^2 - \|\mathbf{x}\|_2^2)$ and a “hard” test function $f_2(\mathbf{x}) = \exp(\sin(20\|\mathbf{x}\|_2)^2 - \|\mathbf{x}\|_2^2)$. The “hard” test function is more variable and will hence be more difficult to approximate (see Fig. 2). One realisation of states $\{\mathbf{x}_i\}_{i=1}^n$ generated independently and uniformly over $\mathcal{X} = [-5, 5]^d$ (initially $d = 1$) was used to estimate the $\Pi[f_i]$. We work in an RKHS characterised by tensor

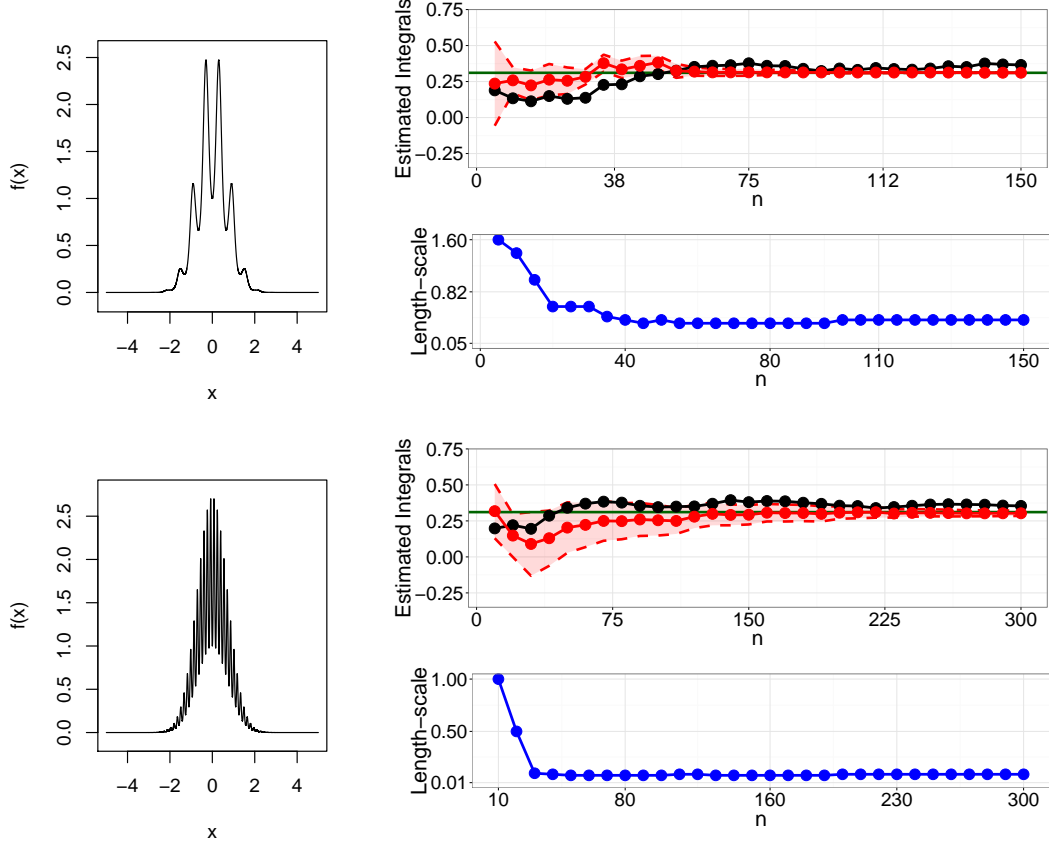


Figure 2: Evaluation of uncertainty quantification provided by empirical Bayes (EB). *Left:* The test functions f_1 (top), f_2 (bottom) in $d = 1$ dimension. *Right:* Solutions provided by Monte Carlo (MC; black) and Bayesian MC (BMC; red), for one typical realisation. BMC was based on a Matérn kernel of order $\alpha = 7/2$. 95% credible regions are shown for BMC and the green horizontal line gives the true value of the integral. The blue curve gives the corresponding lengthscale parameter selected by EB.

products of Matérn kernels

$$k_\alpha(\mathbf{x}, \mathbf{x}') = \lambda^2 \prod_{i=1}^d \frac{2^{1-\alpha}}{\Gamma(\alpha)} \left(\frac{\sqrt{2\alpha}|x_i - x'_i|}{\sigma^2} \right)^\alpha K_\alpha \left(\frac{\sqrt{2\alpha}|x_i - x'_i|}{\sigma^2} \right),$$

where K_α is the modified Bessel function of the second kind. Closed-form kernel means exist in this case for $\alpha = p + 1/2$ whenever $p \in \mathbb{N}$. In this set-up, EB was used to select the length-scale parameter $\sigma \in (0, \infty)$ of the kernel, while magnitude parameter was initially fixed at $\lambda = 1$ and the smoothness parameter $\alpha = 7/2$. Note that all test functions will be in the space \mathcal{H}_α for any $\alpha > 0$ and there is a degree of arbitrariness in this choice of prior.

Results are shown in Fig. 2. Error-bars are used to denote the 95% posterior credible regions for the value of the integral and we also display the values $\hat{\sigma}_n$ selected for σ by EB at each sample size n . The values for $\hat{\sigma}_n$ appear to converge rapidly as $n \rightarrow \infty$; this is encouraging but we emphasise

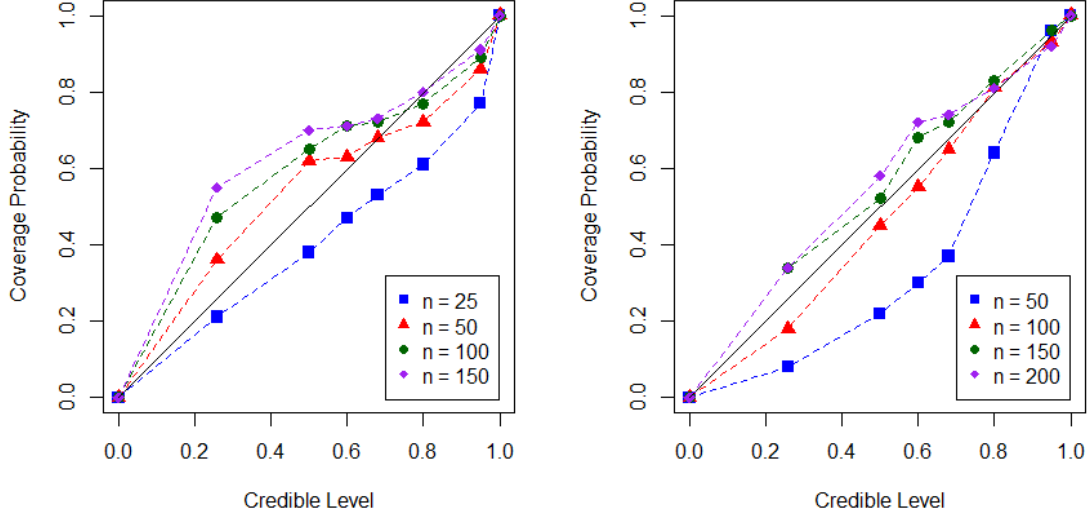


Figure 3: Evaluation of uncertainty quantification provided by empirical Bayes. Coverage frequencies (computed from 100 realisations) were compared against notional $100(1-\alpha)\%$ Bayesian credible regions for varying level α and number of observations n . The upper-left quadrant represents conservative credible intervals whilst the lower-right quadrant represents over-confident intervals. *Left*: “Easy” test function f_1 . *Right*: “Hard” test function f_2 .

that we do not provide theoretical guarantees for EB in this work. On the negative side, over-confidence is possible at small values of n . Indeed, the BQ posterior will be over-confident under EB, since in the absence of evidence to the contrary, EB selects large values for σ that correspond to more regular functions; this is most evident in the “hard” case. Results for varying dimension d are provided in Supplement D.1. Optimising over both (σ, λ) is possible, but this is a noticeably harder problem and requires a larger number of integrand evaluations to avoid over-confident estimation. Extended results for this case can be found also be found in Supplement D.1.

Next we computed coverage frequencies for $100(1-\gamma)\%$ credible regions. For each sample size n , the process was repeated over many realisations of the states $\{\mathbf{x}_i\}_{i=1}^n$, shown in Fig. 3. It may be seen that (for n large enough) the uncertainty quantification provided by EB is over cautious for the easier function f_1 , whilst being well-calibrated for the more complicated functions such as f_2 . As expected, we observed that the coverage was over-confident for small values of n . Performance was subsequently investigated for varying smoothness α and dimension d , with results shown in Supplement D.1.

Finally, to understand whether theoretical results on asymptotic behaviour are realised in practice, we note (in the absence of EB) that the variance $\mathbb{V}_n[\Pi[f]]$ is independent of the integrand and may be plotted as a function of the number n of data. Results in Supplement D.2 demonstrate that theoretical rates are observed in practice; indeed convergence appears to be slightly faster than the upper bounds we provided. At large values of n , numerical instability is observed; this is a well-known computational issue and further commentary is reserved for Supplement C.3.

The results on test functions provided in this section illustrate the extent to which uncertainty

quantification is possible using a BQ approach. In particular, for our examples, EB calibration provided reasonable frequentist coverage if the number n of samples was not too small.

5.2 Case Studies

For the remainder we explore possible roles for BMCMC and BQMC in statistical applications. Four case studies, carefully chosen to highlight both the strengths and the weaknesses of probabilistic integration are presented. Brief critiques of each study are contained below, the full details of which can be found in Supplement E.

5.2.1 Case Study #1: Model Selection via Thermodynamic Integration

Consider the problem of selecting a single best model among a set $\{\mathcal{M}_1, \dots, \mathcal{M}_M\}$, based on data \mathbf{y} assumed to arise from a true model in this set. The Bayesian solution, assuming a uniform prior over models, is to select the *maximum a posteriori* (MAP) model \mathcal{M}_k where $k := \arg \max_{i \in \{1, \dots, M\}} p(\mathcal{M}_i | \mathbf{y})$. We focus on the case with uniform prior on models $p(\mathcal{M}_i) = 1/M$, and this problem hence reduces to finding the largest marginal likelihood $p_i = p(\mathbf{y} | \mathcal{M}_i)$. These terms are usually intractable integrals over the parameters $\boldsymbol{\theta}_i$ associated with model \mathcal{M}_i . One widely-used approach to model selection is to estimate each p_i in turn, say by \hat{p}_i , then to take the maximum of the \hat{p}_i over $i \in \{1, \dots, M\}$. In particular, thermodynamic integration is an effective approach to approximation of marginal likelihoods p_i for individual models (Gelman and Meng, 1998; Friel and Pettitt, 2008).

In many contemporary applications the MAP model is not well-identified, for example in variable selection where there are very many candidate models. Then, the MAP becomes sensitive to numerical error in the \hat{p}_i , since an incorrect model \mathcal{M}_i , $i \neq k$ can be assigned an overly large value of \hat{p}_i due to numerical error, in which case it could be selected in place of the true MAP model. Below we explore the potential to exploit probabilistic integration to surmount this problem.

Thermodynamic Integration To simplify notation below we consider computation of a single p_i and suppress dependence on the index i corresponding to model \mathcal{M}_i . Denote the parameter space by Θ . For $t \in [0, 1]$ (an *inverse temperature*) define the *power posterior* Π_t , a distribution over Θ with density $\pi_t(\boldsymbol{\theta}) \propto p(\mathbf{y} | \boldsymbol{\theta})^t p(\boldsymbol{\theta})$. The thermodynamic identity is formulated as a double integral:

$$\log p(\mathbf{y}) = \int_0^1 dt \int_{\Theta} \log p(\mathbf{y} | \boldsymbol{\theta}) \Pi_t[d\boldsymbol{\theta}].$$

The thermodynamic integral can be re-expressed as

$$\log p(\mathbf{y}) = \int_0^1 g(t) dt, \quad g(t) = \int_{\Theta} f(\boldsymbol{\theta}) \Pi_t[d\boldsymbol{\theta}],$$

where $f(\boldsymbol{\theta}) = \log p(\mathbf{y} | \boldsymbol{\theta})$ and therefore $g(t) = \Pi_t[\log p(\mathbf{y} | \boldsymbol{\theta})]$. Standard practice is to discretise the outer integral and estimate the inner integral using MCMC: Letting $0 = t_1 < \dots < t_m = 1$ denote a fixed *temperature schedule*, we thus have (e.g. using the trapezium rule)

$$\log p(\mathbf{y}) \approx \sum_{i=2}^m (t_i - t_{i-1}) \frac{\hat{g}_i + \hat{g}_{i-1}}{2}, \quad \hat{g}_i = \frac{1}{n} \sum_{j=1}^n \log p(\mathbf{y} | \boldsymbol{\theta}_{i,j}), \quad (7)$$

where $\{\boldsymbol{\theta}_{i,j}\}_{j=1}^n$ are MCMC samples from Π_{t_i} . Several improvements have been proposed, including the use of higher-order numerical quadrature for the outer integral (Friel et al., 2014; Hug et al., 2015) and the use of control variates for the inner integral (Oates et al., 2016b,c). To date, probabilistic integration has not been explored in this setting.

Probabilistic Thermodynamic Integration Our proposal is to apply probabilistic integration to both the inner and outer integrals. This is instructive, since nested integrals are prone to propagation and accumulation of numerical error. Several features of the method are highlighted:

Transfer Learning: In the probabilistic approach to thermodynamic integration, the two integrands f and g are each assigned prior probability models. For the inner integral we assign a prior $f \sim \mathcal{N}(0, k_f)$. Our data here are the $nm \times 1$ vector \mathbf{f} where $f_{(i-1)n+j} = f(\boldsymbol{\theta}_{i,j})$. For estimating g_i with BQ we have m times as much data as for the MC estimator \hat{g}_i , in Eqn. 7, which makes use of only n function evaluations. Here, information transfer across temperatures is made possible by the explicit model for f underpinning the probabilistic integrator.

In the posterior, $\mathbf{g} = [g(t_1), \dots, g(t_T)]$ is a Gaussian random vector with $\mathbf{g}|\mathbf{f} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where the mean and covariance are defined by

$$\begin{aligned}\mu_a &= \Pi_{t_a}[\mathbf{k}_f(\cdot, X)]\mathbf{K}_f^{-1}\mathbf{f}, \\ \Sigma_{a,b} &= \Pi_{t_a}\Pi_{t_b}[\mathbf{k}_f(\cdot, \cdot)] - \Pi_{t_a}[\mathbf{k}_f(\cdot, X)]\mathbf{K}_f^{-1}\Pi_{t_b}[\mathbf{k}_f(X, \cdot)],\end{aligned}$$

where $X = \{\boldsymbol{\theta}_{i,j}\}_{j=1}^n$ and \mathbf{K}_f is an $nm \times nm$ kernel matrix defined by k_f .

Inclusion of Prior Information: For the outer integral, it is known that discretisation error can be substantial; Friel et al. (2014) proposed a second-order correction to the trapezium rule to mitigate this bias, while Hug et al. (2015) pursued the use of Simpson’s rule. Attacking this problem from the probabilistic perspective, we do not want to place a default prior on $g(t)$, since it is known from extensive empirical work that $g(t)$ will vary more at smaller values of t . Indeed the rule-of-thumb $t_i = (i/m)^5$ is commonly used to inform the choice of quadrature states in accordance with this observation (Calderhead and Girolami, 2009). We would like to encode this information into our prior. To do this, we proceed with an importance sampling step $\log p(\mathbf{y}) = \int_0^1 g(t)dt = \int_0^1 h(t)\pi(t)dt$. The rule-of-thumb implies that taking an importance distribution Π with density $\pi(t) \propto 1/(\epsilon + 5t^{4/5})$ for some small $\epsilon > 0$, which renders the function $h = g/\pi$ approximately stationary (made precise in Supplement E.1). A stationary GP prior $h \sim \mathcal{N}(0, k_h)$ on the transformed integrand h provides an encoding of this prior knowledge that was used for experiments.

Propagation of Uncertainty: Under this construction, in the posterior $\log p(\mathbf{y})$ is Gaussian with mean and covariance defined as

$$\begin{aligned}\mathbb{E}_n[\log p(\mathbf{y})] &= \Pi[\mathbf{k}_h(\cdot, T)]\mathbf{K}_h^{-1}\boldsymbol{\mu} \\ \mathbb{V}_n[\log p(\mathbf{y})] &= \underbrace{\Pi\Pi[\mathbf{k}_h(\cdot, \cdot)] - \Pi[\mathbf{k}_h(\cdot, T)]\mathbf{K}_h^{-1}\Pi[\mathbf{k}_h(T, \cdot)]}_{(*)} + \underbrace{\Pi[\mathbf{k}_h(\cdot, T)]\mathbf{K}_h^{-1}\boldsymbol{\Sigma}\mathbf{K}_h^{-1}\Pi[\mathbf{k}_h(T, \cdot)]}_{(**)},\end{aligned}$$

where $T = \{t_i\}_{i=1}^m$ and \mathbf{K}_h is an $m \times m$ kernel matrix defined by k_h . The term $(*)$ arises from BQ on the outer integral, while the term $(**)$ arises from propagating numerical uncertainty from the inner integral through to the outer integral.

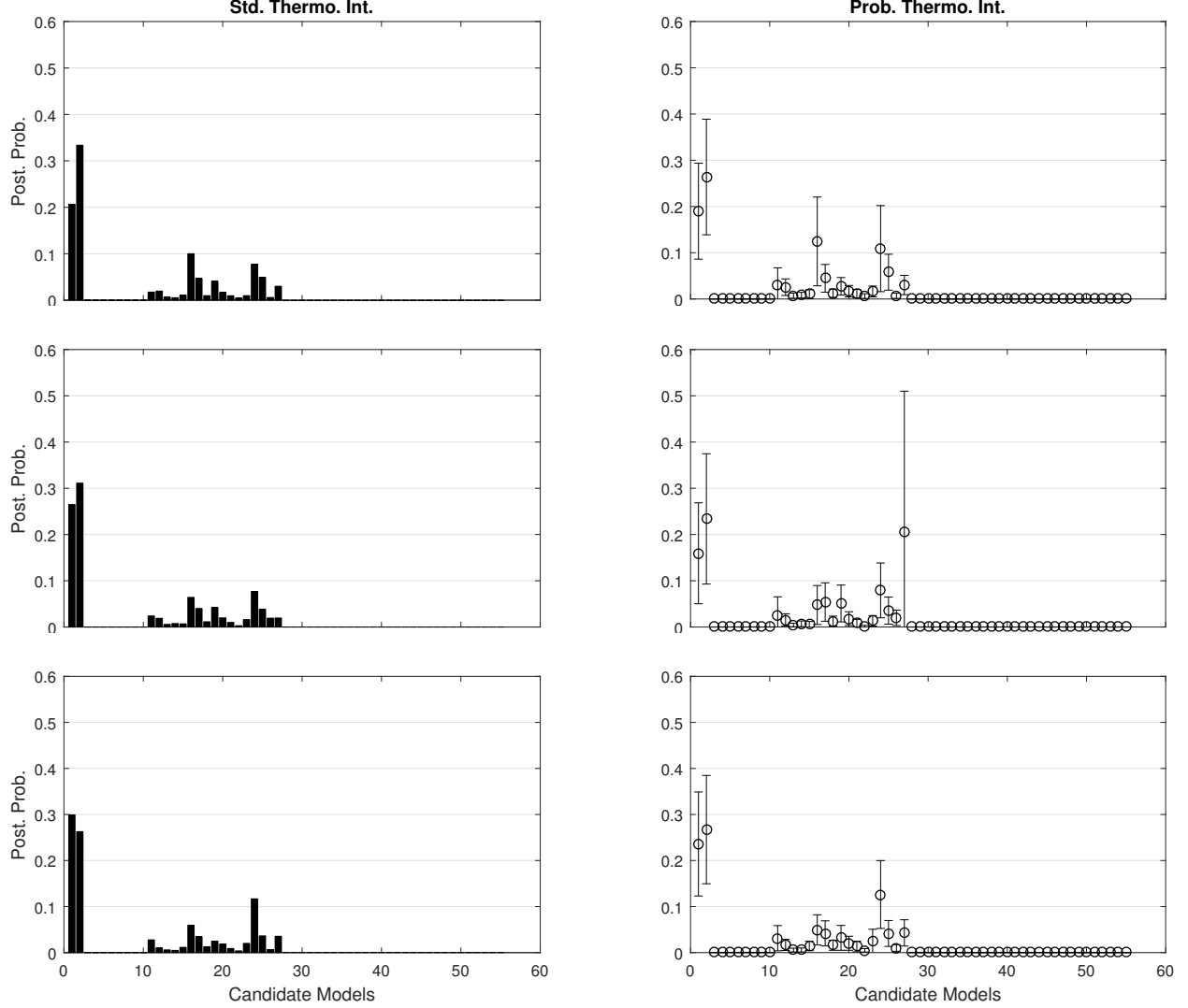


Figure 4: Probabilistic thermodynamic integration; illustration on variable selection for logistic regression (the true model was \mathcal{M}_1). Standard and probabilistic thermodynamic integration were used to approximate marginal likelihoods and, hence, the posterior over models. Each row represents an independent realisation of the MCMC sampler, while the data \mathbf{y} were fixed throughout. *Left*: Standard Monte Carlo, where point estimates for marginal likelihood were assumed to have no associated numerical error. *Right*: Probabilistic integration, where a model for numerical error on each integral was propagated through into the posterior over models. This was performed using BMCMC and thermodynamic integration with BQ weights. The probabilistic approach produces a “probability distribution over a probability distribution”, where the numerical uncertainty is modelled on top of the usual uncertainty associated with model selection.

Simulation Study An experiment was conducted to elicit the MAP model from a collection of 56 candidate logistic regression models in a variable selection setting. This could be achieved in many ways; our aim was not to compare accuracy of point estimates, but rather to explore the

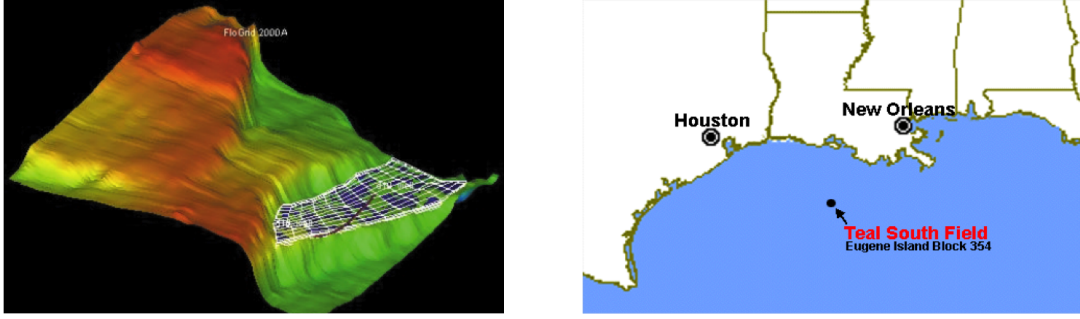


Figure 5: *Left:* Computer model for the Teal South oil field. Simulation of this model requires significant computational resources. This renders any statistical analysis challenging due to the small number of data which can be obtained. *Right:* Location of the oil field.

probability model that, unlike in standard methods, is provided by probabilistic integration. Full details are provided in Supplement E.1.

Results are shown in Fig. 4. Here we compared approximations to the model posterior obtained using the standard method versus the probabilistic method, over 3 realisations of the MCMC (the data \mathbf{y} were fixed). We make some observations: (i) The probabilistic approach produces a second-order probability distribution, where the numerical uncertainty is modelled on top of the usual uncertainty associated with model selection. (ii) The computation associated with BQ required less time, in total, than the time taken to obtain MCMC samples. (iii) The same model was not always selected as the MAP when numerical error was ignored and depended on the MCMC random seed. In contrast, under the probabilistic approach, either \mathcal{M}_1 or \mathcal{M}_2 could feasibly be the MAP under any of the MCMC realisations, up to numerical uncertainty. (iv) The middle row of Fig. 4 shows a large posterior uncertainty over the marginal likelihood for \mathcal{M}_{27} . This suggests more computational effort should be expended on this particular integral. (v) The posterior variance was dominated by uncertainty due to discretisation error in the outer integral, rather than the inner integral. This suggests that numerical uncertainty could be reduced by allocating more computational resources to the outer integral rather than the inner integral.

This section concludes by noting the quite natural connection between algorithm design and models for numerical error, as exemplified by points (iv) and (v) above.

5.2.2 Case Study #2: Uncertainty Quantification for Computer Experiments

Here we consider an industrial scale computer model for the Teal South oil field (Hajizadeh et al., 2011), situated off the coast of New Orleans (depicted in Fig. 5). Conditional on field data, posterior inference was facilitated using state-of-the-art MCMC methodology (Lan et al., 2016). Oil reservoir models are generally challenging for standard MCMC methods for several reasons. First, simulating from those models can be prohibitively computationally expensive, making the cost of individual MCMC samples from a few minutes to several hours. Second, the posterior distribution will often exhibit strongly non-linear concentration of measure. Here we computed statistics of interest using BMCMC, where the uncertainty quantification afforded by BQ aims to enable valid inferences in the presence of a relatively small number of MCMC samples. Full details are provided in Supplement E.2.

Quantification of the uncertainty associated with predictions is a major topic of ongoing research

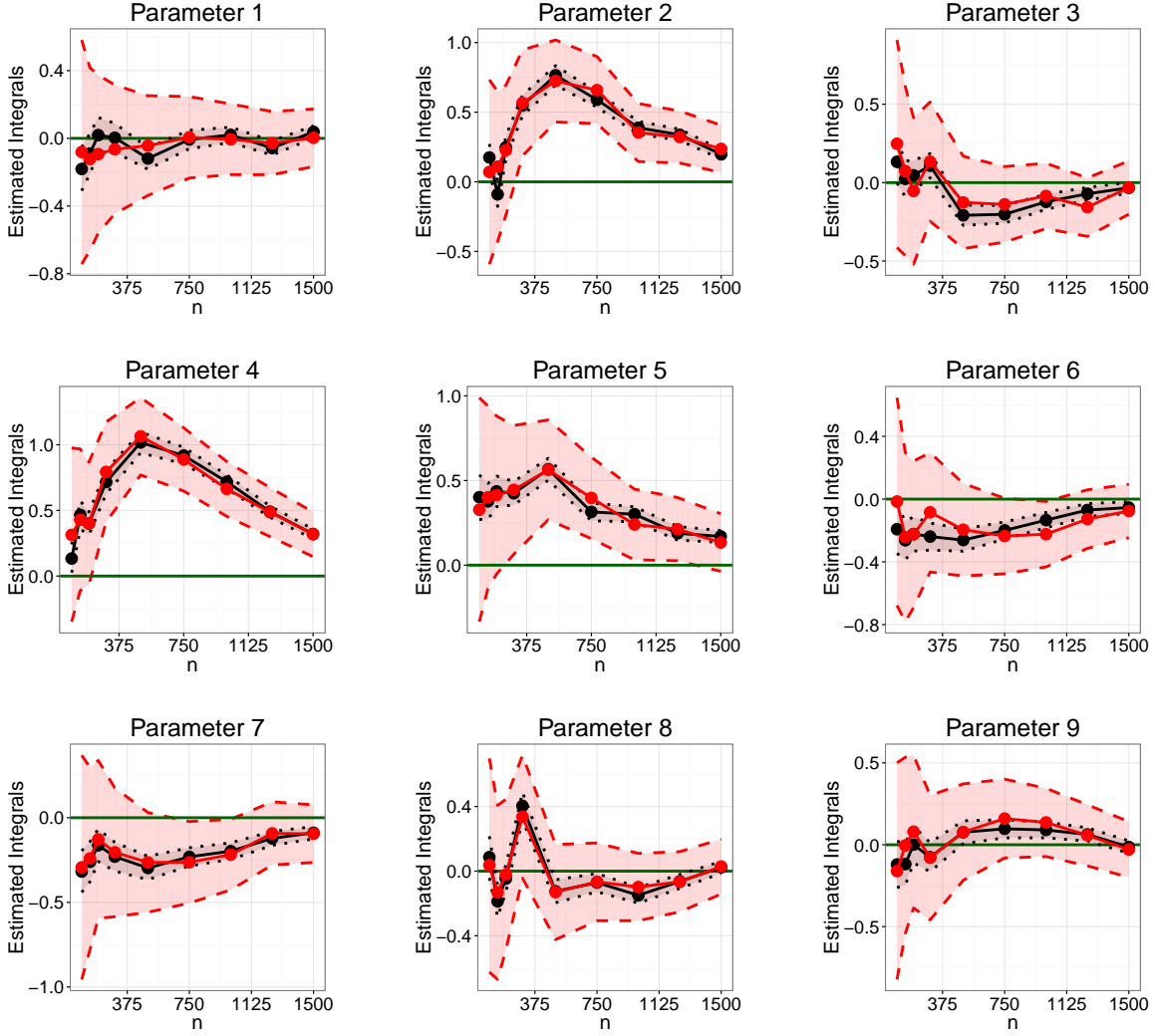


Figure 6: Numerical estimation of posterior parameter means for Teal South (centered around the true values). The green line gives the exact value of the integral. The MCMC (black line) and BMC MC point estimates (red line) provided similar performance. The MCMC 95% confidence intervals are based on estimated asymptotic variance (black dotted lines) tend to be over confident whereas with the BMC MC 95% credible intervals (red dotted lines) provide a more honest quantification of numerical uncertainty.

in this field (Mohamed et al., 2010; Hajizadeh et al., 2011; Park et al., 2013) due to the economic consequences associated with inaccurate predictions of quantities such as future oil production rate. A probabilistic model for numerical error in integrals associated with prediction could provide a more honest assessment that factors in additional sources of uncertainty.

The particular integrals that we considered are posterior means for each model parameter, and we compared against an empirical benchmark obtained with brute force from a long MCMC run. BMC MC was employed with a Matérn $\alpha = 3/2$ kernel whose lengthscale-parameter was selected

using EB. Estimates for posterior means were obtained using both standard MCMC and BMCMC, shown in Fig. 6. For this example the posterior distribution afforded by BMCMC provides sensible uncertainty quantification for parameters 1, 3, 6-9, but was over-confident for parameters 2, 4, 5. The point accuracy of the BMCMC estimator matched that of the standard MCMC estimator. The lack of faster convergence for BMCMC appears to be due to inaccurate estimation of the kernel mean and we conjecture that alternative exact approaches, such as Oates et al. (2016b), may provide improved performance in this context. However, standard confidence intervals obtained from the CLT for MCMC with a plug-in estimate for the asymptotic variance were over-confident for parameters 2-9.

5.2.3 Case Study #3: High-Dimensional Random Effects

Our aim here is to explore whether the more flexible functional representations afforded by weighted combinations of Hilbert spaces enable probabilistic integration in more challenging higher-dimensional settings. The focus was BQMC, but the methodology could be applied to other probabilistic integrators.

Weighted Spaces The formulation of high (and infinite) -dimensional QMC results requires a construction known as a *weighted* Hilbert space. These spaces, defined below, are motivated by the observation that many integrands encountered in applications seem to vary more in lower dimensional projections compared to higher dimensional projections. Our presentation below follows Sec. 2.5.4 and 12.2 of Dick and Pillichshammer (2010), but the idea goes back at least to Wahba (1990, Chap. 10).

As usual with QMC, we work in $\mathcal{X} = [0, 1]^d$ and Π uniform over \mathcal{X} . Let $\mathcal{I} = \{1, 2, \dots, d\}$. For each subset $u \subseteq \mathcal{I}$, define a weight $\gamma_u \in (0, \infty)$ and denote the collection of all weights by $\gamma = \{\gamma_u\}_{u \subseteq \mathcal{I}}$. Consider the space \mathcal{H}_γ of functions of the form $f(\mathbf{x}) = \sum_{u \subseteq \mathcal{I}} f_u(\mathbf{x}_u)$, where f_u belongs to an RKHS \mathcal{H}_u with reproducing kernel k_u and \mathbf{x}_u denotes the components of \mathbf{x} that are indexed by $u \subseteq \mathcal{I}$. This construction is not restrictive, since any function can be written in this form by considering only $u = \mathcal{I}$. We turn \mathcal{H}_γ into a Hilbert space by defining an inner product $\langle f, g \rangle_\gamma := \sum_{u \subseteq \mathcal{I}} \gamma_u^{-1} \langle f_u, g_u \rangle_u$ where $\gamma = \{\gamma_u : u \subseteq \mathcal{I}\}$. Constructed in this way, \mathcal{H}_γ is an RKHS with reproducing kernel $k_\gamma(\mathbf{x}, \mathbf{x}') = \sum_{u \subseteq \mathcal{I}} \gamma_u k_u(\mathbf{x}, \mathbf{x}')$. Intuitively, the weights γ_u can be taken to be small whenever the function f does not depend heavily on the $|u|$ -way interaction of the states \mathbf{x}_u . Thus, most of the γ_u will be small for a function f that is effectively low-dimensional. A measure of the dimensionality of the function is given by $\sum_{u \subseteq \mathcal{I}} \gamma_u$; in an extreme case d could even be infinite provided that this sum remains bounded (Dick et al., 2013).

The (canonical) *weighted* Sobolev space of dominating mixed smoothness $\mathcal{S}_{\alpha, \gamma}$ is defined by taking each of the component spaces to be \mathcal{S}_α . In finite dimensions $d < \infty$, BQMC rules based on a higher-order digital net attain optimal WCE rates $O(n^{-\alpha+\epsilon})$ for this RKHS; see Supplement E.3 for full details.

Semi-Parametric Random Effects Regression For illustration we considered generalised linear models, and focus on a Poisson semi-parametric random effects regression model studied by

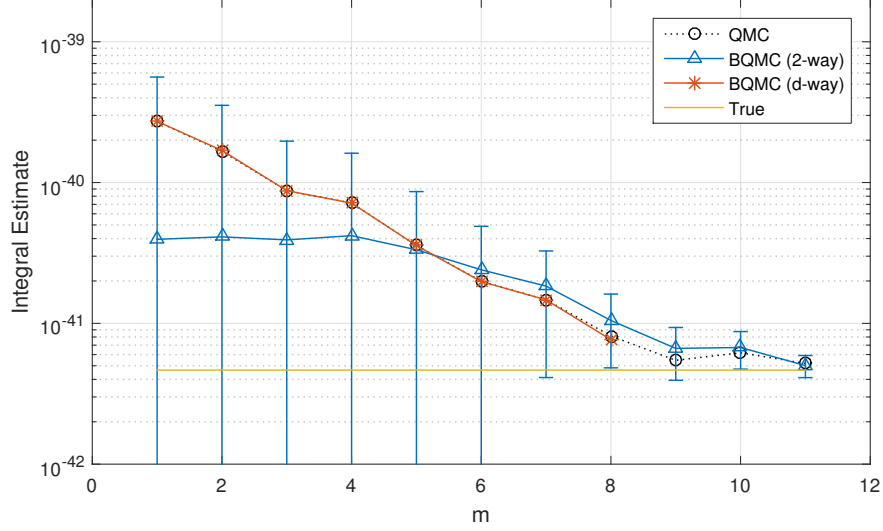


Figure 7: Application to semi-parametric random effects regression in $d = 50$ dimensions, based on $n = 2^m$ samples from a higher-order digital net. [Error bars show 95% credible regions. To improve visibility results are shown on the log-scale; error bars are symmetric on the linear scale. A brute-force QMC estimate was used to approximate the true value of the integral $p(\mathbf{y}|\boldsymbol{\beta})$ where $\boldsymbol{\beta} = (0, 1, 1)$ was the data-generating value of the parameter.]

Kuo et al. (2008, Example 2). The context is inference for the parameters $\boldsymbol{\beta}$ of the following model

$$\begin{aligned} Y_j | \lambda_j &\sim \text{Po}(\lambda_j) \\ \log(\lambda_j) &= \beta_0 + \beta_1 z_{1,j} + \beta_2 z_{2,j} + u_1 \phi_1(z_{2,j}) + \cdots + u_d \phi_d(z_{2,j}) \\ u_j &\sim N(0, \tau^{-1}) \text{ independent.} \end{aligned}$$

Here $z_{1,j} \in \{0, 1\}$, $z_{2,j} \in (0, 1)$ and $\phi_j(z) = [z - \kappa_j]_+$ where $\kappa_j \in (0, 1)$ are pre-determined knots. We took $d = 50$ equally spaced knots in $[\min \mathbf{z}_2, \max \mathbf{z}_2]$. Inference for $\boldsymbol{\beta}$ requires multiple evaluations of the observed data likelihood $p(\mathbf{y}|\boldsymbol{\beta}) = \int_{\mathbb{R}^d} p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}) p(\mathbf{u}) d\mathbf{u}$ and therefore is a natural candidate for probabilistic integration methods, in order to propagate the cumulative uncertainty of estimating multiple numerical integrals into the posterior distribution $p(\boldsymbol{\beta}|\mathbf{y})$.

In order to transform this integration problem to the unit cube we perform the change of variables $x_j = \Phi^{-1}(u_j)$ so that we wish to evaluate $p(\mathbf{y}|\boldsymbol{\beta}) = \int_{[0,1]^d} p(\mathbf{y}|\boldsymbol{\beta}, \Phi^{-1}(\mathbf{x})) d\mathbf{x}$. Here $\Phi^{-1}(\mathbf{x})$ denotes the standard Gaussian inverse CDF applied to each component of \mathbf{x} . Probabilistic integration here proceeds under the hypothesis that the integrand $f(\mathbf{x}) = p(\mathbf{y}|\boldsymbol{\beta}, \Phi^{-1}(\mathbf{x}))$ belongs to (or at least can be well approximated by functions in) $\mathcal{S}_{\alpha, \gamma}$ for some smoothness parameter α and some weights γ . Intuitively, the integrand $f(\mathbf{x})$ is such that an increase in the value of x_j at the knot κ_j can be compensated for by a decrease in the value of x_{j+1} at a neighbouring knot κ_{j+1} , but not by changing values of \mathbf{x} at more remote knots. Therefore we expect $f(\mathbf{x})$ to exhibit strong individual and pairwise dependence on the x_j , but expect higher-order dependency to be much weaker. This motivates the weighted space assumption. Sinescu et al. (2012) provides theoretical analysis for the choice of weights γ . Here, weights γ of *order two* were used; $\gamma_u = 1$ for $|u| \leq d_{\max}$, $d_{\max} = 2$, $\gamma_u = 0$ otherwise, which corresponds to an assumption of low order interaction terms (though f can still depend on all d of its arguments). Full details are provided in Supplement E.3.

Results in Fig. 7 showed that the 95% posterior credible regions more-or-less cover the truth for this problem, suggesting that the uncertainty estimates are appropriate. On the negative side, the BQMC method does not encode non-negativity of the integrand and, consequently, some posterior mass is placed on negative values for the integral, which is not meaningful. To understand the effect of the weighted space construction here, we compared against BQMC with d -way interactions ($u \in \{\emptyset, \mathcal{I}\}$). An interesting observation was that these point estimates closely followed those produced by QMC.

5.2.4 Case Study #4: Spherical Integration for Computer Graphics

Probabilistic integration methods can be defined on arbitrary manifolds, with formulations on non-Euclidean spaces suggested as far back as Diaconis (1988) and recently exploited in the context of computer graphics (Brouillat et al., 2009; Marques et al., 2015). This forms the setting for our final case study.

Global Illumination Integrals Below we formulate and analyse BQMC on the d -sphere $\mathbb{S}^d = \{\mathbf{x} = (x_1, \dots, x_{d+1}) \in \mathbb{R}^{d+1} : \|\mathbf{x}\|_2 = 1\}$ in order to estimate integrals of the form $\Pi[f] = \int_{\mathbb{S}^d} f d\Pi$, where Π is the spherical measure (i.e. uniform over \mathbb{S}^d with $\int_{\mathbb{S}^d} d\Pi = 1$).

Probabilistic integration is applied to compute global illumination integrals used in the rendering of surfaces (Pharr and Humphreys, 2004), and we therefore focus on the case where $d = 2$ and the measure Π is uniform over \mathbb{S}^2 . The problem of global illumination occurs in the synthesis of photo-realistic images of virtual scenes (e.g. a view of a lake). Uncertainty quantification is motivated by inverse global illumination problems (e.g. Yu et al., 1999), where the task is to make inferences from noisy observation of an object via computer-based image synthesis; a measure of numerical uncertainty could naturally be propagated in such problems. Below, to limit scope, we restrict attention to uncertainty quantification in the forward problem.

The models involved in global illumination are based on three main factors: a geometric model for the objects present in the scene, a model for the reflectivity of the surface of each object and a description of the light sources (provided by an *environment map* as depicted in Fig. 8). The light emitted from the environment will interact with objects in the scene through reflection. This can be formulated as an illumination integral of the form below:

$$L_o(\boldsymbol{\omega}_o) = L_e(\boldsymbol{\omega}_o) + \int_{\mathbb{S}^2} L_i(\boldsymbol{\omega}_i) \rho(\boldsymbol{\omega}_i, \boldsymbol{\omega}_o) [\boldsymbol{\omega}_i \cdot \mathbf{n}]_+ \Pi(d\boldsymbol{\omega}_i). \quad (8)$$

Here $L_o(\boldsymbol{\omega}_o)$ is the *outgoing radiance*, i.e. the outgoing light in the direction $\boldsymbol{\omega}_o$. $L_e(\boldsymbol{\omega}_o)$ represents the amount of light emitted by the object itself (which we will assume to be known) and $L_i(\boldsymbol{\omega}_i)$ is the light hitting the object from direction $\boldsymbol{\omega}_i$. The term $\rho(\boldsymbol{\omega}_i, \boldsymbol{\omega}_o)$ is the *bidirectional reflectance distribution* function (BRDF), which models the fraction of light arriving at the surface point from direction $\boldsymbol{\omega}_i$ and being reflected towards direction $\boldsymbol{\omega}_o$. Here \mathbf{n} is a unit vector normal to the surface of the object. Our investigation is motivated by strong empirical results for BQMC in this context obtained by Marques et al. (2015).

In order to assess the performance of BQMC we consider a typical illumination integration problem based on the California lake environment map shown in Fig. 8. The goal here is to compute intensities for each of the three RGB colour channels corresponding to observing a virtual object from a fixed direction $\boldsymbol{\omega}_o$. We consider the case of an object directly facing the camera ($\boldsymbol{\omega}_o = \mathbf{n}$). For the BRDF we took $\rho(\boldsymbol{\omega}_i, \boldsymbol{\omega}_o) = (2\pi)^{-1} \exp(\boldsymbol{\omega}_i \cdot \boldsymbol{\omega}_o - 1)$. The integral in Eqn. 8 was viewed here

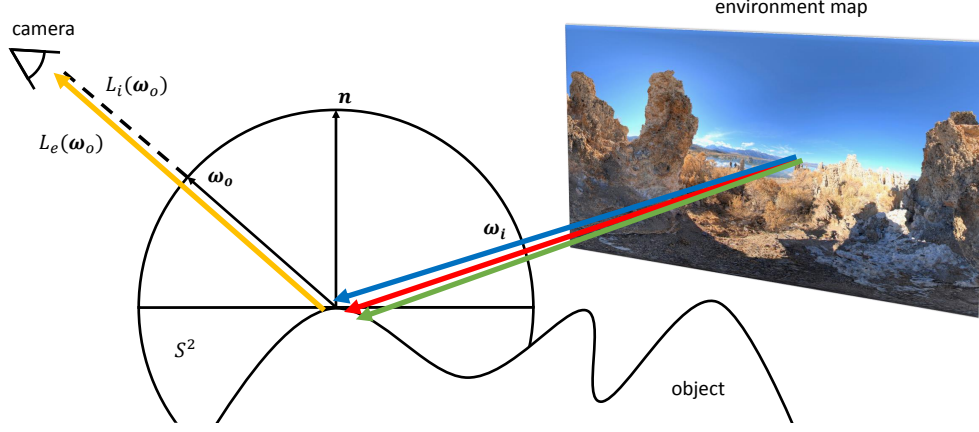


Figure 8: Application to illumination integrals in computer graphics. The cartoon features the California lake environment map that was used in our experiments.

as an integral with respect to a uniform measure Π with integrand $f(\omega_i) = L_i(\omega_i)\rho(\omega_i, \omega_o)[\omega_i \cdot \omega_o]_+$ assumed to be in a Sobolev space of low smoothness. In contrast, Marques et al. (2015) viewed Eqn. 8 as an integral with respect to $\pi(\omega_i) \propto \rho(\omega_i, \omega_o)$ and coupled this with a Gaussian kernel restricted to the hemisphere. The approach that we propose has two advantages; (i) it provides a closed-form expression for the kernel mean, (ii) a rougher kernel may be more appropriate in the context of illumination integrals, as pointed out by Brouillat et al. (2009). The specific function space that we consider is the Sobolev space $\mathcal{H}_\alpha(\mathbb{S}^d)$ for $\alpha = 3/2$, formally defined in Supplement E.4.

Results Both BMC and BQMC were tested on this example. To ensure fair comparison, identical kernels were taken as the basis for both methods. BQMC was employed using a *spherical t-design* (Bondarenko et al., 2013). It can be shown that for BQMC $\|\hat{\Pi}_{\text{BQMC}} - \Pi\|_{\mathcal{H}^*} = O(n^{-3/4})$ when this point set is used (see Supplement E.4).

Fig. 9 shows performance in RGB-space. For this particular test function, the BQMC point estimate was almost identical to the QMC estimate at all values of n . Empirical results reported by Marques et al. (2015), based on Gaussian kernels, showed a RMSE rate of $O(n^{-0.72})$, which is similar to the theoretical $O(n^{-3/4})$ rate for BQMC. Overall, both BMC and BQMC provided sensible quantification of uncertainty for the value of the integral at all values of n that were considered.

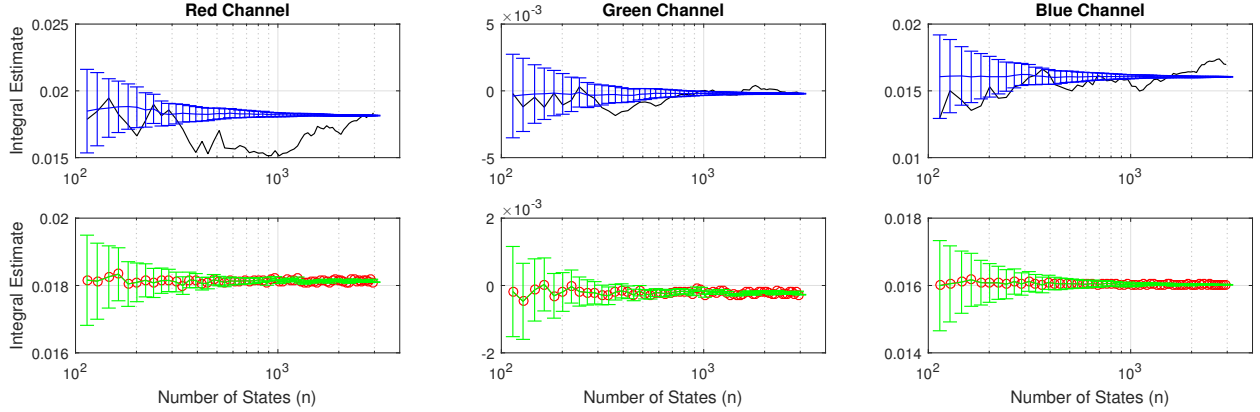


Figure 9: Probabilistic integration over the sphere was employed to estimate the RGB colour intensities for the California lake environment. [Error bars for BMC (blue) and BQMC (green) represent 95% credible intervals. MC estimates (black) and QMC estimates (red) are shown for reference.]

6 Conclusion

The increasing sophistication of complex computational models, of which numerical integration is one component, demands an improved understanding of how numerical error accumulates and propagates through sequential computation. In (now common) settings where integrands are computationally intensive, or very many numerical integrals are required, effective methods are required that make full use of information available about the problem at hand. This is clearly evidenced by the recent success of QMC methods, which leverage the smoothness properties of integrands. Probabilistic Numerics puts the statistician in centre stage and aims to *model* the unknown numerical error. This approach was eloquently summarised by Kadane (1985), who proposed the following vision for the future of computation:

“Statistics can be thought of as a set of tools used in making decisions and inferences in the face of uncertainty. Algorithms typically operate in such an environment. Perhaps then, statisticians might join the teams of scholars addressing algorithmic issues.”

This paper explored probabilistic integration from the perspective of the statistician. Our results highlight both the advantages and disadvantages of such an approach. On the positive side, the general methodology described a unified framework in which existing high-performance MCMC and QMC methods can be associated with a probability distribution that captures the extent of their numerical error. Posterior contraction rates were established for the first time and form an important, fundamental and novel contribution of this work. On the negative side, there remain many substantial open questions and challenges surrounding probabilistic integration, in terms of philosophical foundations, theoretical analysis and practical application. These are discussed below:

Philosophy There are several issues concerning the uncertainty which one models with probabilistic numerical methods. In the specific case of integration, there is not yet a consensus on how one should interpret uncertainty in the estimation of a constant; is it as simple as “epistemic uncertainty” due to the finite number of evaluations of the integrands? If so, whose epistemic

uncertainty is being modelled? The discussion that accompanied Kong et al. (2003) is notable by the lack of consensus among statisticians.

An up-to-date discussion of these points is provided in Hennig et al. (2015). There it was argued that the uncertainty being modelled is that of a hypothetical agent “that we get to design”. That is, the statistician selects priors and loss functions for the agent so that it best achieves the statistician’s own goals. These goals typically involve a combination of relatively black-box behaviour, to perform well on a diverse range of problems, and a low computational overhead. Interpretation of the posterior is then more subtle than for subjective inference and many of the points of contention for objective inference also appear in this framework.

Methodology The question of which part of the numerical method should be modelled is open to debate. In this paper, the integrand f is considered to be uncertain while the distribution Π is considered to be known. However, one could alternatively suppose that both f and Π are unknown, pursued in Oates et al. (2016d).

Irrespective of where uncertainty is introduced, the scope to design numerical methods that specifically target numerical error has close links to statistical decision theory (Diaconis, 1988). For example, Briol et al. (2015) consider the choice of quadrature states as a problem in statistical experimental design, providing an iterative approach that is proven to asymptotically minimise posterior variance over the unknown value of the integral. This also leads to the question of which statistical tools are most appropriate, and this is likely to be a question with no overall answer, but which will be application dependent and will depend on the philosophical tendencies of the user.

Theory For probabilistic integration, further theoretical work is required. Our results proved several desirable properties of the behaviour of posterior contraction, but did not address coverage at finite sample size, nor the interaction of coverage with the EB method for kernel parameter selection. A particularly important question, addressed in Kanagawa et al. (2016), is the behaviour of BQ when the integrand does not belong to the posited RKHS.

Prior Specification A broad discussion is required on what prior information should be included, and what information should be ignored. Indeed, practical considerations essentially always demand that some aspects of prior information are ignored. Competing computational, statistical and philosophical considerations are all in play and must be balanced.

For example, the RKHS framework that we studied in this paper has the advantage of providing a flexible way to encode prior knowledge about the integrand, allowing to specify properties such as smoothness, periodicity, non-stationarity and effective low-dimension. On the other hand, several important properties, including positivity, are less easily encoded. For BQ, the possibility for importance sampling (Eqn. 6) has an element of arbitrariness that appears to preclude the pursuit of a default prior. At a more fundamental level, the constraint that $f \in L^2(\Pi)$ can be critiqued as physically irrelevant (in physical problems where f is meaningful, f^2 can be nonsensical), but this is intrinsic to the Hilbert space approach.

Even within the RKHS framework, there is the issue that integrands f will usually be part of infinitely many RKHS. Selecting an appropriate RKHS is arguably the central open challenge for QMC research at present. From a practical perspective, elicitation of priors over infinite-dimensional spaces is a hard problem. An adequate choice of prior can be very informative for the numerical scheme and can significantly improve the convergence rates of the method. Methods for

choosing the kernel automatically would be useful here (e.g. Duvenaud, 2014), but would need to be considered against their suitability for providing correct uncertainty quantification.

The list above is not meant to be exhaustive, but highlights the many areas of research that are yet to be explored. In conclusion, it is clear to us that there is an important role for statisticians to play in the future development of modern numerical analysis.

References

- Bach, F. (2015). On the equivalence between quadrature rules and random features. *arXiv:1502.06800*.
- Bardenet, R. and Hardy, A. (2016). Monte Carlo with Determinantal Point Processes. *arXiv:1605.00361*.
- Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, New York.
- Bissiri, P., Holmes, C. and Walker, S. (2016). A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, in press.
- Bondarenko, A., Radchenko, D. and Viazovska, M. (2013). Optimal asymptotic bounds for spherical designs. *Ann. of Math.*, 178(2):443–452.
- Brauchart, J., Saff, E., Sloan, I. H. and Womersley, R. (2014). QMC designs: Optimal order quasi Monte Carlo integration schemes on the sphere. *Math. Comp.*, 83:2821–2851.
- Briol, F.-X., Oates, C. J., Girolami, M. and Osborne, M. A. (2015). Frank-Wolfe Bayesian quadrature: Probabilistic integration with theoretical guarantees. In *Adv. Neur. Inf. Proc. Sys.*, 1162–1170.
- Brouillat, J., Bouville, C., Loos, B., Hansen, C. and Bouatouch, K. (2009). A Bayesian Monte Carlo approach to global illumination. *Comp. Graph. Forum*, 28(8):2315–2329.
- Calderhead, B. and Girolami, M. (2009). Estimating Bayes factors via thermodynamic integration and population MCMC. *Comput. Statist. Data Anal.*, 53(12):4028–4045.
- Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2016). Stan: A probabilistic programming language. *J Statist. Software*, in press.
- Cockayne, J., Oates, C. J., Sullivan, T. and Girolami, M. (2016). Probabilistic Numerical Methods for PDE-constrained Bayesian Inverse Problems. *Proc. 36th I. Workshop Bayesian Infer. Max. Entropy Methods Sci. Eng.*, to appear.
- Conrad, P., Girolami, M., Särkkä, S., Stuart, A., and Zygalakis, K. (2016). Probability measures for numerical solutions of differential equations. *Stat. Comput.*, to appear.
- Dashti, M. and Stuart, A. (2016). The Bayesian approach to inverse problems. In *Handbook of Uncertainty Quantification*.

- Davis, P. J. and Rabinowitz, P. (2007). *Methods of numerical integration*. Courier Corporation.
- Deylon, B. and Portier, F. (2016). *Integral approximation by kernel smoothing*. *Bernoulli*, 22(4): 2177–2208.
- Diaconis, P. (1988). Bayesian numerical analysis. *Statist. Decis. Theory Rel. Top. IV*, 163–175.
- Dick, J. and Pillichshammer, F. (2010). *Digital nets and sequences - discrepancy theory and quasi-Monte Carlo integration*. Cambridge University Press.
- Dick, J., Kuo, F. Y. and Sloan, I. H. (2013). High-dimensional integration: The quasi-Monte Carlo way. *Acta Numer.*, 22:133–288.
- Duvenaud, D. (2014). *Automatic model construction with Gaussian processes*. PhD thesis, University of Cambridge.
- Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.
- Eftang, J. L. and Stamm, B. (2012). Parameter multi-domain “hp” empirical interpolation. *Int. J. Numer. Methods Eng.*, 90(4):412–428.
- Fasshauer, G., Hickernell, F. and Woźniakowski, H. (2012). On dimension-independent rates of convergence for function approximation with Gaussian kernels. *SIAM J. Numer. Anal.*, 50(1): 247–271.
- Friel, N. and Pettitt, A. (2008). Marginal likelihood estimation via power posteriors. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 70(3):589–607.
- Friel, N. Hurn, M. and Wyse, J. 2014. Improving power posterior estimation of statistical evidence. *Stat. Comput.*, 24(5):709–723.
- Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.*, 13(2):163–185.
- Gerber, M. and Chopin, N. (2015). Sequential quasi-Monte Carlo. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 77(3):509–579.
- Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 73(2):123–214.
- Gunter, T., Garnett, R., Osborne, M., Hennig, P. and Roberts, S. (2014). Sampling for inference in probabilistic models with fast Bayesian quadrature. In *Adv. Neur. Inf. Proc. Sys.*, 2789–2797.
- Hajizadeh, Y., Christie, M., and Demyanov, V. (2011). Ant colony optimization for history matching and uncertainty quantification of reservoir models. *J. Petrol. Sci. Eng.*, 77(1):78–92.
- Hennig, P. (2015). Probabilistic interpretation of linear solvers. *SIAM J. Optim.*, 25(1):234–260.
- Hennig, P. and Kiefel, M. (2013). Quasi-Newton methods: A new direction. *J. Mach. Learn. Res.*, 14:843–865.

- Hennig, P., Osborne, M. A. and Girolami, M. (2015). Probabilistic numerics and uncertainty in computations. *J. Roy. Soc. A*, 471(2179).
- Hickernell, F. J. (1998). A generalized discrepancy and quadrature error bound. *Math. Comp.*, 67(221):299–322.
- Hickernell, F. J., Lemieux, C., and Owen, A. B. (2005). Control variates for quasi-Monte Carlo. *Statist. Sci.*, 20(1):1–31.
- Hug, S., Schwarzfischer, M., Hasenauer, J., Marr, C., and Theis, F. J. (2015). An adaptive scheduling scheme for calculating Bayes factors with thermodynamic integration using Simpson’s rule. *Stat. Comp.*, 1–15.
- Huszar, F. and Duvenaud, D. (2012). Optimally-weighted herding is Bayesian quadrature. In *Uncertain. Artif. Intell.*, 377–385.
- Kadane, J. B. (1985). Parallel and sequential computation: A statistician’s view. *J. Complexity*, 1:256–263.
- Kadane, J. B. and Waskowski, G. W. (1985). Average case ϵ -complexity in computer science: A Bayesian view. In *Bayesian Statist.*, 2:361–374.
- Kanagawa, M., Sriperumbudur, B. K. and Fukumizu, K. (2016). Convergence guarantees for kernel-based quadrature rules in misspecified settings. In *Adv. Neur. Inf. Proc. Sys.*, to appear.
- Kong, A., McCullagh, P., Meng, X.-L., Nicolae, D., and Tan, Z. (2003). A theory of statistical models for Monte Carlo integration. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 65(3):585–618.
- Kristoffersen, S. (2013). The empirical interpolation method. Master’s thesis, Department of Mathematical Sciences, Norwegian University of Science and Technology.
- Kuo, F. Y., Dunsmuir, W. T. M., Sloan, I. H., Wand, M. P. and Womersley, R. S. (2008). Quasi-Monte Carlo for highly structured generalised response models. *Methodol. Comput. Appl. Probab.*, 10(2):239–275.
- Lan, S., Bui-Thanh, T., Christie, M., and Girolami, M. (2016). Emulation of higher-order tensors in manifold Monte Carlo methods for Bayesian inverse problems. *J. Comput. Phys.*, 308:81–101.
- Maday, Y., Nguyen, N. C., Patera, A. T., and Pau, G. S. H. (2009). A general, multipurpose interpolation procedure: the magic points. *Commun. Pure Appl. Anal.*, 8:383–404.
- Mahsereci, M. and Hennig, P. (2015). Probabilistic line searches for stochastic optimization. In *Adv. Neur. Inf. Proc. Sys.*, 181–189.
- Marques, R., Bouville, C., Ribardiere, M., Santos, L. P. and Bouatouch, K. (2013). A spherical Gaussian framework for Bayesian Monte Carlo rendering of glossy surfaces. In *IEEE Trans. Vis. and Comp. Graph.*, 19(10):1619–1632.
- Marques, R., Bouville, C., Santos, L. and Bouatouch, K. (2015). Efficient quadrature rules for illumination integrals: from quasi Monte Carlo to Bayesian Monte Carlo. *Synth. Lect. Comput. Graph. Animation*, 7(2):1–92.

- Minka, T. (2000). Deriving quadrature rules from Gaussian processes. Technical report, Statistics Department, Carnegie Mellon University.
- Mizielinski, M., Roberts, M., Vidale, P., Schiemann, R., Demory, M., Strachan, J., et al. (2014). High-resolution global climate modelling: the upscale project, a large-simulation campaign. *Geosci. Model Dev.*, 7(4):1629–1640.
- Mohamed, L., Christie, M. and Demyanov, V. (2010). Comparison of stochastic sampling algorithms for uncertainty quantification. *SPE Journal*, 15:31–38.
- Mosbach, S. and Turner, A. G. (2009). A quantitative probabilistic investigation into the accumulation of rounding errors in numerical ODE solution. *Comput. Math. Appl.*, 57(7):1157–1167.
- Novak, E. and Woźniakowski, H. (2008). *Tractability of multivariate problems, volume I: Linear information*. EMS Publishing House, EMS Tracts in Mathematics 6.
- Novak, E. and Woźniakowski, H. (2010). *Tractability of multivariate problems, volume II : Standard information for functionals*. EMS Publishing House, EMS Tracts in Mathematics 12.
- Nuyens, D. and Cools, R. (2006). Fast component-by-component construction, a reprise for different kernels. In *Monte Carlo and Quasi-Monte Carlo Methods 2004*, 373–387. Springer.
- Oates, C. J., Cockayne, J., Briol, F.-X. and Girolami, M. (2016). Convergence rates for a class of estimators based on Stein’s identity. *arXiv:1603.03220*.
- Oates, C. J., Girolami, M., and Chopin, N. (2016). Control functionals for Monte Carlo integration. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, in press.
- Oates, C. J., Papamarkou, T., and Girolami, M. (2016). The controlled thermodynamic integral for Bayesian model comparison. *J. Amer. Statist. Assoc.*, in press.
- Oates, C. J., Briol, F. X., and Girolami, M. (2016). Probabilistic integration and intractable distributions. *arXiv:1606.06841*.
- O’Hagan, A. (1991). Bayes–Hermite quadrature. *J. Statist. Plann. Inference*, 29:245–260.
- O’Hagan, A. (1992). Some Bayesian numerical analysis. *Bayesian Statist.*, 4:345–363.
- Osborne, M. A. (2010). *Bayesian Gaussian processes for sequential prediction, optimisation and quadrature*. PhD thesis, University of Oxford.
- Osborne, M. A., Duvenaud, D., Garnett, R., Rasmussen, C. E., Roberts, S. and Ghahramani, Z. (2012). Active learning of model evidence using Bayesian quadrature. In *Adv. Neur. Inf. Proc. Sys.*, 46–54.
- Owhadi, H. (2016). Multi-grid with rough coefficients and multiresolution operator decomposition from hierarchical information games. *SIAM Rev.*, in press.
- Park, H., Scheidt, C., Fenwick, D., Boucher, A., and Caers, J. (2013). History matching and uncertainty quantification of facies models with multiple geological interpretations. *Comput. Geosci.*, 17(4):609–621.

- Pharr, M. and Humphreys, G. (2004). *Physically based rendering: From theory to implementation*. Morgan Kaufmann Publishers Inc.
- Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press.
- Rasmussen, C. E. and Ghahramani, Z. (2002). Bayesian Monte Carlo. In *Adv. Neur. Inf. Proc. Sys.*, 489–496.
- Ritter, K. (2000). *Average-case analysis of numerical problems*. Springer-Verlag Berlin Heidelberg.
- Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.
- Särkka, S., Hartikainen, J., Svensson, L., and Sandblom, F. (2016). On the relation between Gaussian process quadratures and sigma-point methods, *J. Adv. Inf. Fusion*, in press.
- Schaback, R. (1995). Error estimates and condition numbers for radial basis function interpolation. *Adv. Comput. Math.*, 3:251–264.
- Schober, M., Duvenaud, D., and Hennig, P. (2014). Probabilistic ODE solvers with Runge-Kutta means. In *Adv. Neur. Inf. Proc. Sys.*, 739–747.
- Schölkopf, B. and Smola, A. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press.
- Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and de Freitas, N. (2015). Taking the human out of the loop: A review of Bayesian optimization. In *Proc. IEEE*, 104(1):148–175.
- Sickel, W. and Ullrich, T. (2009). Tensor products of Sobolev–Besov spaces and applications to approximation from the hyperbolic cross. *J. Approx. Theory*, 161(2):748–786.
- Sinescu, V., Kuo, F. Y. and Sloan, I. H. (2012). On the choice of weights in a function space for quasi-Monte Carlo methods for a class of generalised response models in statistics. In *Monte Carlo and Quasi-Monte Carlo Methods*.
- Skilling, J. (1991). Bayesian solution of ordinary differential equations. *Max. Entropy Bayesian Methods, Seattle*.
- Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A Hilbert space embedding for distributions. In *Proc. 18th I. Conf. Algorithmic Learn. Theory*, 13–31.
- Sobol, I. (1993). Sensitivity estimates for non linear mathematical models. *Math. Mod. Comput. Exp.*, 1:407–414.
- Sommariva, A. and Vianello, M. (2006). Numerical cubature on scattered data by radial basis functions. *Computing*, 76(3-4):295–310.
- Song, L. (2008). *Learning via Hilbert space embedding of distributions*. PhD thesis, School of Information Technologies, University of Sydney.

- Sozzetti, A., Giacobbe, P., Lattanzi, M. G., Micela, G., Morbidelli, R. and Tinetti, G. (2013). Astrometric detection of giant planets around nearby M dwarfs: the gaia potential. *Monthly Not. Roy. Astronom. Soc.*.
- Stein, M. (2008). *Interpolation of Spatial Data - Some Theory for Kriging*. Springer Science & Business Media.
- Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.
- Suldin, A. B. (1959). Wiener measure and its applications to approximation methods. *I. Izvestiya Vysshikh Uchebnykh Zavedenii. Matematika*, 6145–158.
- Szabó, B., van der Vaart, A. and van Zanten, J. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.*, 43(4):1391–1428.
- Traub, J. F., Woźniakowski, H. and Wasilkowski, G. W. (1988). Information-based complexity. Academic Press.
- Triebel, H. (1992). *Theory of function spaces II*. Birkhäuser Basel.
- van Der Vaart, A. and van Zanten, H. (2011). Information rates of nonparametric Gaussian process methods. *J. Mach. Learn. Res.*, 12:2095–2119.
- Wahba, G. (1990). *Spline models for observational data*, volume 59. SIAM.
- Wendland, H. (2005). *Scattered data approximation*. Cambridge University Press.
- Yang, Y. and Dunson, D. B. (2013). Bayesian manifold regression. *arXiv:1305.0617*.
- Yu, Y., Debevec, P., Malik, J. and Hawkins, T. (1999). Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *Proc. Ann. Conf. Comput. Graph. Int. Tech.*, 215–224.

Supplement

This supplement provides complete proofs for theoretical results, extended numerical results and full details to reproduce the experiments presented in the paper *Probabilistic Integration: A Role for Statisticians in Numerical Analysis?* by Briol, Oates, Girolami, Osborne & Sejdinovic. The structure of this document is as follows:

A Proof of Theoretical Results	1
B Generalisation of Theoretical Results	5
C Scalability, Stability and Kernel Means	6
C.1 Scalability in the Number of States	6
C.2 Scalability in Dimension	7
C.3 Regularisation and Noisy Observations	8
C.4 Intractable Kernel Means	10
D Additional Numerical Results	12
D.1 Empirical Bayes Calibration	12
D.2 Empirical Rates of Posterior Contraction	12
E Supplemental Information for Case Studies	15
E.1 Case Study #1	15
E.2 Case Study #2	17
E.3 Case Study #3	18
E.4 Case Study #4	19

A Proof of Theoretical Results

Full technical details are provided to establish the theoretical results from the main text:

Proof of Prop. 1. For a prior $\mathcal{GP}(m, k)$ and data $\{(\mathbf{x}_i, f_i)\}_{i=1}^n$, standard conjugacy results for GPs lead to the posterior $\mathbb{P}_n = \mathcal{GP}(m_n, k_n)$ over $L^2(\Pi)$, where

$$\begin{aligned} m_n(\mathbf{x}) &= m(\mathbf{x}) + \mathbf{k}(\mathbf{x}, X) \mathbf{K}^{-1}(\mathbf{f} - \mathbf{m}) \\ k_n(\mathbf{x}, \mathbf{x}') &= k(\mathbf{x}, \mathbf{x}') - \mathbf{k}(\mathbf{x}, X) \mathbf{K}^{-1} \mathbf{k}(X, \mathbf{x}'), \end{aligned}$$

see Chap. 2 of Rasmussen and Williams (2006). Then repeated application of Fubini's theorem produces

$$\begin{aligned} \mathbb{E}_n[\Pi[f]] &= \mathbb{E}_n \left[\int_{\mathcal{X}} f \, d\Pi \right] = \int_{\mathcal{X}} m_n \, d\Pi \\ \mathbb{V}_n[\Pi[f]] &= \int_{\mathcal{F}} \left[\int_{\mathcal{X}} f \, d\Pi - \int_{\mathcal{X}} m_n \, d\Pi \right]^2 \mathbb{P}_n[df] \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} \int_{\mathcal{F}} [f(\mathbf{x}) - m_n(\mathbf{x})][f(\mathbf{x}') - m_n(\mathbf{x}')] \mathbb{P}_n[df] \Pi(d\mathbf{x}) \Pi(d\mathbf{x}') \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} k_n(\mathbf{x}, \mathbf{x}') \Pi(d\mathbf{x}) \Pi(d\mathbf{x}'). \end{aligned}$$

The proof is completed by substituting the expressions for m_n and k_n into these two equations. \square

Proof of Prop. 2. From Eqn. 5 in the main text

$$\|\hat{\Pi} - \Pi\|_{\mathcal{H}^*} \leq \|\mu(\hat{\Pi}) - \mu(\Pi)\|_{\mathcal{H}}.$$

For the converse inequality, consider the specific integrand $f = \mu(\hat{\Pi}) - \mu(\Pi)$. Then from the supremum definition of the dual norm $\|\cdot\|_{\mathcal{H}^*}$:

$$\|\hat{\Pi} - \Pi\|_{\mathcal{H}^*} \geq \frac{|\hat{\Pi}[f] - \Pi[f]|}{\|f\|_{\mathcal{H}}}.$$

Now we use the reproducing property:

$$\begin{aligned} \frac{|\hat{\Pi}[f] - \Pi[f]|}{\|f\|_{\mathcal{H}}} &= \frac{|\langle f, \mu(\hat{\Pi}) - \mu(\Pi) \rangle_{\mathcal{H}}|}{\|f\|_{\mathcal{H}}} \\ &= \frac{\|\mu(\hat{\Pi}) - \mu(\Pi)\|_{\mathcal{H}}^2}{\|\mu(\hat{\Pi}) - \mu(\Pi)\|_{\mathcal{H}}} = \|\mu(\hat{\Pi}) - \mu(\Pi)\|_{\mathcal{H}}. \end{aligned}$$

This completes the proof. \square

Proof of Prop. 3. Combining Prop. 2 with direct calculation gives that

$$\begin{aligned} \|\hat{\Pi} - \Pi\|_{\mathcal{H}^*}^2 &= \|\mu(\hat{\Pi}) - \mu(\Pi)\|_{\mathcal{H}}^2 \\ &= \sum_{i,j=1}^n w_i w_j k(\mathbf{x}_i, \mathbf{x}_j) - 2 \sum_{i=1}^n w_i \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}_i) \Pi(d\mathbf{x}) + \int_{\mathcal{X}} \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}') \Pi(d\mathbf{x}) \Pi(d\mathbf{x}') \\ &= \mathbf{w}^T \mathbf{K} \mathbf{w} - 2 \mathbf{w}^T \Pi[\mathbf{k}(X, \cdot)] + \Pi \Pi[k(\cdot, \cdot)] \end{aligned}$$

as required. \square

The following lemma shows that probabilistic integrators provide a point estimate that is *at least as good* as their non-probabilistic versions. Consistency of the non-probabilistic integrators is therefore a sufficient condition to prove consistency of their probabilistic counterparts.

Lemma 2 (Bayesian re-weighting). *Let $f \in \mathcal{H}$. Consider the quadrature rule $\hat{\Pi}[f] = \sum_{i=1}^n w_i f(\mathbf{x}_i)$ and the corresponding BQ rule $\hat{\Pi}_{BQ}[f] = \sum_{i=1}^n w_i^{BQ} f(\mathbf{x}_i)$. Then $\|\hat{\Pi}_{BQ} - \Pi\|_{\mathcal{H}^*} \leq \|\hat{\Pi} - \Pi\|_{\mathcal{H}^*}$.*

Proof. This is immediate from Prop. 3, which shows that the BQ weights w_i^{BQ} are an optimal choice for the space \mathcal{H} . \square

The convergence rate of quadrature rules can be shown to be at least as good as the corresponding functional approximation rates in $L^2(\Pi)$. This is summarised as follows:

Lemma 3 (Regression bound). *Let $f \in \mathcal{H}$ and fix states $\{\mathbf{x}_i\}_{i=1}^n \in \mathcal{X}$. Then we have $|\Pi[f] - \hat{\Pi}_{BQ}[f]| \leq \|f - m_n\|_2$.*

Proof. This is an application of Jensen's inequality:

$$\begin{aligned} |\Pi[f] - \hat{\Pi}_{\text{BQ}}[f]|^2 &= \left(\int_{\mathcal{X}} f \, d\Pi - \int_{\mathcal{X}} m_n \, d\Pi \right)^2 \\ &\leq \int_{\mathcal{X}} (f - m_n)^2 \, d\Pi = \|f - m_n\|_2^2, \end{aligned}$$

as required. \square

Note that this regression bound is not sharp in general (Ritter, 2000, Prop. II.4) and as a consequence Thm. 1 below is not quite optimal, as discussed in the main text.

Lemmas 2 and 3 refer to the point estimators provided by BQ rules. However, we also aim to quantify the change in probability mass as the number of samples increases:

Lemma 4 (BQ contraction). *Assume $f \in \mathcal{H}$. Suppose that $\|\hat{\Pi}_{\text{BQ}} - \Pi\|_{\mathcal{H}^*} \leq \delta_n$ where $\delta_n \rightarrow 0$ as $n \rightarrow \infty$. Define $I_D = [\Pi[f] - D, \Pi[f] + D]$ to be an interval of radius D centred on the true value of the integral. Then $\mathbb{P}_n[I_D^c]$, the posterior mass on $I_D^c = \mathbb{R} \setminus I_D$, vanishes at the rate $\mathbb{P}_n[I_D^c] = O(\exp(-(D^2/2)\delta_n^{-2}))$.*

Proof. Assume without loss of generality that $D < \infty$. The posterior distribution over $\Pi[f]$ is Gaussian with mean m_n and variance v_n . Since $v_n = \|\hat{\Pi}_{\text{BQ}} - \Pi\|_{\mathcal{H}^*}^2$ we have $v_n \leq \delta_n^2$. Now the posterior probability mass on I_D^c is given by $\mathbb{P}_n[I_D^c] = \int_{I_D^c} \phi(r|m_n, v_n) dr$, where $\phi(r|m_n, v_n)$ is the p.d.f. of the $\mathcal{N}(m_n, v_n)$ distribution. From the definition of D we get the upper bound

$$\begin{aligned} \mathbb{P}_n[I_D^c] &\leq \int_{-\infty}^{\Pi[f]-D} \phi(r|m_n, v_n) dr + \int_{\Pi[f]+D}^{\infty} \phi(r|m_n, v_n) dr \\ &= 1 + \Phi\left(\underbrace{\frac{\Pi[f] - m_n}{\sqrt{v_n}}}_{(*)} - \frac{D}{\sqrt{v_n}}\right) - \Phi\left(\underbrace{\frac{\Pi[f] - m_n}{\sqrt{v_n}}}_{(*)} + \frac{D}{\sqrt{v_n}}\right). \end{aligned}$$

From the definition of the WCE we have that the terms $(*)$ are bounded by $\|f\|_{\mathcal{H}} < \infty$, so that asymptotically as $\delta_n \rightarrow 0$ we have

$$\begin{aligned} \mathbb{P}_n[I_D^c] &\lesssim 1 + \Phi(-D/\sqrt{v_n}) - \Phi(D/\sqrt{v_n}) \\ &\lesssim 1 + \Phi(-D/\delta_n) - \Phi(D/\delta_n) \\ &\lesssim \text{erfc}(D/\sqrt{2}\delta_n). \end{aligned}$$

The result follows from the fact that $\text{erfc}(x) \lesssim \exp(-x^2/2)$ for x sufficiently small. \square

This result demonstrates that the posterior distribution is well-behaved; probability mass concentrates on the true solution as n increases. Hence, if our prior is well calibrated (see Sec. 4.1), the posterior distribution provides valid uncertainty quantification over the solution of the integral as a result of performing a finite number n of integrand evaluations.

Define the *fill distance* of the set $X = \{\mathbf{x}_i\}_{i=1}^n$ as

$$h_X = \sup_{\mathbf{x} \in \mathcal{X}} \min_{i=1, \dots, n} \|\mathbf{x} - \mathbf{x}_i\|_2.$$

As $n \rightarrow \infty$ the scaling of the fill distance is described by the following:

Lemma 5. Let $g : [0, \infty) \rightarrow [0, \infty)$ be continuous, monotone increasing, and satisfy $g(0) = 0$ and $\lim_{x \downarrow 0} g(x) \exp(x^{-3d}) = \infty$. Suppose further $\mathcal{X} = [0, 1]^d$, Π has a density π that is bounded away from zero on \mathcal{X} , and $X = \{\mathbf{x}_i\}_{i=1}^n$ are samples from an uniformly ergodic Markov chain targeting Π . Then we have $\mathbb{E}_X[g(h_X)] = O(g(n^{-1/d+\epsilon}))$ where $\epsilon > 0$ can be arbitrarily small.

Proof. A special case of Lemma 2, Oates et al. (2016a). \square

Proof of Thm. 1. Initially consider fixed states $X = \{\mathbf{x}_i\}_{i=1}^n$ (i.e. fixing the random seed) and $\mathcal{H} = \mathcal{H}_\alpha$. From a now standard result in functional approximation due to Wu and Schaback (1993), see also Wendland (2005, Thm. 11.13) there exists $C > 0$ and $h_0 > 0$ such that, for all $\mathbf{x} \in \mathcal{X}$ and $h_X < h_0$,

$$|f(\mathbf{x}) - m_n(\mathbf{x})| \leq Ch_X^\alpha \|f\|_{\mathcal{H}}.$$

(For other kernels, alternative bounds are well-known; Wendland, 2005, Table 11.1). We augment X with a finite number of states $Y = \{\mathbf{y}_i\}_{i=1}^m$ to ensure that $h_{X \cup Y} < h_0$ always holds. Then from the regression bound (Lemma 3),

$$\begin{aligned} |\hat{\Pi}_{\text{BMCMC}}[f] - \Pi[f]| &\leq \|f - m_n\|_2 = \left(\int_{\mathcal{X}} (f(\mathbf{x}) - m_n(\mathbf{x}))^2 \Pi(d\mathbf{x}) \right)^{1/2} \\ &\leq \left(\int_{\mathcal{X}} (Ch_{X \cup Y}^\alpha \|f\|_{\mathcal{H}})^2 \Pi(d\mathbf{x}) \right)^{1/2} = Ch_{X \cup Y}^\alpha \|f\|_{\mathcal{H}}. \end{aligned}$$

It follows that $\|\hat{\Pi}_{\text{BMCMC}} - \Pi\|_{\mathcal{H}_\alpha^*} \leq Ch_{X \cup Y}^\alpha$. Now, taking an expectation \mathbb{E}_X over the sample path $X = \{\mathbf{x}_i\}_{i=1}^n$ of the Markov chain, we have that

$$\mathbb{E}_X \|\hat{\Pi}_{\text{BMCMC}} - \Pi\|_{\mathcal{H}_\alpha^*} \leq C \mathbb{E}_X h_{X \cup Y}^\alpha \leq C \mathbb{E}_X h_X^\alpha. \quad (9)$$

From Lemma 5 above, we have a scaling relationship such that, for $h_{X \cup Y} < h_0$, we have $\mathbb{E}_X h_X^\alpha = O(n^{-\alpha/d+\epsilon})$ for $\epsilon > 0$ arbitrarily small. From Markov's inequality, convergence in mean implies convergence in probability and thus, using Eqn. 9, we have $\|\hat{\Pi}_{\text{BMCMC}} - \Pi\|_{\mathcal{H}_\alpha^*} = O_P(n^{-\alpha/d+\epsilon})$. This completes the proof for $\mathcal{H} = \mathcal{H}_\alpha$. More generally, if \mathcal{H} is norm-equivalent to \mathcal{H}_α then the result follows from the fact that $\|\hat{\Pi}_{\text{BMCMC}} - \Pi\|_{\mathcal{H}^*} \leq \lambda \|\hat{\Pi}_{\text{BMCMC}} - \Pi\|_{\mathcal{H}_\alpha^*}$ for some $\lambda > 0$. \square

Proof of Thm. 2. From Theorem 15.21 of Dick and Pillichshammer (2010), which assumes $\alpha \geq 2$, $\alpha \in \mathbb{N}$, the QMC rule $\hat{\Pi}_{\text{QMC}}$ based on a higher-order digital $(t, \alpha, 1, \alpha m \times m, d)$ net over \mathbb{Z}_b for some prime b satisfies

$$\|\hat{\Pi}_{\text{QMC}} - \Pi\|_{\mathcal{H}^*} \leq C_{d,\alpha} \frac{(\log n)^{d\alpha}}{n^\alpha} = O(n^{-\alpha+\epsilon})$$

for \mathcal{S}_α the Sobolev space of dominating mixed smoothness order α , where $C_{d,\alpha} > 0$ is a constant that depends only on d and α (but not on n). The result follows immediately from Bayesian re-weighting (Lemma 2) and norm equivalence. The contraction rate is obtained by applying Lemma 4. \square

B Generalisation of Theoretical Results

In this section we extend the theoretical results provided for B(MC)MC in Supplement A to RKHS which are not Sobolev spaces. We show convergence rates for these probabilistic integrators for any RKHS in situations where we have access to an upper bound on the $L^\infty(\mathcal{X})$ error:

$$\sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x}) - m_n(\mathbf{x})| \leq Cg(h_X)\|f\|_{\mathcal{H}}, \quad (10)$$

where the role of g can be compared with that of the *power function* in the scattered data approximation literature (see Wendland (2005)[Sec. 11.1] for more details on this function). For example, for \mathcal{H}_α we have $g(h_X) = h_X^\alpha$; this was the basis of our analysis of B(MC)MC. The results presented below hold for B(MC)MC, but can also be obtained for various QMC methods (not presented).

The following is a slight generalisation of Thm. 1:

Theorem 3. *Let $\mathcal{X} = [0, 1]^d$ and let Π admit a density π that is bounded away from zero on \mathcal{X} . Consider $X = \{\mathbf{x}_i\}_{i=1}^n$ to be samples from an uniformly ergodic Markov chain targeting Π . Suppose Eqn. 10 is satisfied for all $h_X < h_0$, for some $h_0 < \infty$, where g is a continuous, monotone increasing function $g : [0, \infty) \rightarrow [0, \infty)$ satisfying $g(0) = 0$ and $\lim_{x \downarrow 0} g(x) \exp(x^{-3d}) = \infty$. Then, $\|\hat{\Pi}_{BMCMC} - \Pi\|_{\mathcal{H}^*} = O_P(g(m^{-1/d+\epsilon}))$ for all $\epsilon > 0$ arbitrarily small.*

Proof. Following the reasoning of the proof of Thm. 1, we obtain that

$$|\hat{\Pi}_{BMCMC}[f] - \Pi[f]| \leq Cg(h_{X \cup Y})\|f\|_{\mathcal{H}},$$

where we have again augmented the data set from X to $X \cup Y$ to ensure $h_{X \cup Y} \leq h_0$. Taking expectations, we then get the following upper bound: $\mathbb{E}_X \|\hat{\Pi}_{BMCMC} - \Pi\|_{\mathcal{H}^*} \leq \mathbb{E}_X [g(h_{X \cup Y})]$. Now, Lemma 5 shows that $\mathbb{E}_X [g(h_{X \cup Y})] = O(g(m^{-1/d+\epsilon}))$, which leads to $\mathbb{E}_X \|\hat{\Pi}_{BMCMC} - \Pi\|_{\mathcal{H}^*} = O(g(m^{-1/d+\epsilon}))$. The result $\|\hat{\Pi}_{BMCMC} - \Pi\|_{\mathcal{H}^*} = O_P(g(m^{-1/d+\epsilon}))$ follows by noting that convergence in expectation implies convergence in probability. \square

The generalisation in Thm. 3 allows us to obtain convergence rates for B(MC)MC (and BQMC) based on three kernels extensively used in the statistics literature:

- *Gaussian* kernel: $k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|_2^2 / 2\sigma^2)$, $\sigma \in (0, \infty)$.
- *Multi-Quadric* (MQ) kernel: $k(\mathbf{x}, \mathbf{y}) = (-1)^{\lceil \beta \rceil} (c^2 + \|\mathbf{x} - \mathbf{y}\|_2^2)^\beta$, $\beta > 0$, $\beta \notin \mathbb{N}$, $c \in \mathbb{R}$.
- *Inverse Multi-Quadric* (Inverse MQ) kernel: $k(\mathbf{x}, \mathbf{y}) = (c^2 + \|\mathbf{x} - \mathbf{y}\|_2^2)^{-\beta}$, $\beta < 0$, $c \in \mathbb{R}$.

Here we present results for the convergence of WCE, but note that Lemma 4 immediately implies specific rates for posterior contraction:

Corollary 1. *Under the hypotheses of Thm. 3, let k be any of the Gaussian, MQ or inverse MQ kernels. Then there exists $c \in (0, \infty)$ such that $\|\hat{\Pi}_{BMCMC} - \Pi\|_{\mathcal{H}^*} = O_P(\exp(-cn^{1/d-\epsilon}))$ for all $\epsilon > 0$ arbitrarily small.*

Proof. From Table 11.1 in Wendland (2005), we obtain upper bounds on the power function for the Gaussian, MQ and inverse MQ kernels. In the case of the Gaussian, this is given by $g_1(h_X) = \exp(-c_1 |\log(h_X)|/h_X) = \exp(-c_1/h_X^{1-\epsilon'})$ for some $c_1 > 0$ and $\epsilon' > 0$ arbitrarily small. For the MQ and inverse MQ kernels this is $g_2(h_X) = \exp(-c_2/h_X)$ for some $c_2 > 0$. We are now interested in

the behaviour of the WCE, which we obtain using Thm. 3. For the Gaussian kernel, this is given by

$$\begin{aligned}\|\hat{\Pi}_{\text{BMCMC}} - \Pi\|_{\mathcal{H}^*} &= O_P(g_1(n^{-1/d+\epsilon})) \\ &= O_P(\exp(-c_1 n^{1/d-\epsilon''})),\end{aligned}$$

where $\epsilon'' > 0$ can be arbitrarily small, whilst for the MQ and inverse MQ we have

$$\begin{aligned}\|\hat{\Pi}_{\text{BMCMC}} - \Pi\|_{\mathcal{H}^*} &= O_P(g_2(n^{-1/d+\epsilon})) \\ &= O_P(\exp(-c_2 n^{1/d-\epsilon})).\end{aligned}$$

This completes the proof. \square

Similar results can be obtained for power kernels, thin-plate splines and compact support kernels using the bounds provided in (Wendland, 2005, Sec. 11). Results for Gaussian kernels on \mathbb{R}^d (a non-compact domain) are provided in Fasshauer et al. (2012).

Finally, we note that all of the theoretical results in the main text, as well as the present section, can be generalised to domains satisfying interior cone conditions. Following Wendland (2005)[Sec. 3.3], a domain $\mathcal{X} \in \mathbb{R}^d$ is said to satisfy an *interior cone condition* if there exists an angle $\theta \in (0, \pi/2)$ and a radius $r > 0$ such that $\forall \mathbf{x} \in \mathcal{X}$, a unit vector $\boldsymbol{\xi}(\mathbf{x})$ exists such that the following cone is contained in \mathcal{X} :

$$C(\mathbf{x}, \boldsymbol{\xi}(\mathbf{x}), \theta, r) := \left\{ \mathbf{x} + \lambda \mathbf{y} : \mathbf{y} \in \mathbb{R}^d, \|\mathbf{y}\|_2 = 1, \mathbf{y}^T \boldsymbol{\xi}(\mathbf{x}) \geq \cos \theta, \lambda \in [0, r] \right\}.$$

This condition essentially excludes domains with pinch-points on the boundaries, i.e. regions with a \prec shape. The main step involved in this generalisation of our results is to establish Lemma 5 for domains that satisfy an interior cone condition. This contribution was made in Oates et al. (2016a).

C Scalability, Stability and Kernel Means

This section is concerned with details relating to the implementation of BQ, including numerical issues of scalability and stability, as well as a novel approach to approximation of kernel means that was employed in case studies #1 and #2.

C.1 Scalability in the Number of States

In situations where f is cheap to evaluate, the naive $O(n^3)$ computational cost associated with kernel matrix inversion renders BQ more computationally intensive relative to the $O(n)$ cost of non-probabilistic (MC)MC and QMC methods. It is then natural to ask whether the uncertainty quantification provided in BQ is worth the increased off-line computational overhead. Below several approaches to reducing the computational overhead of BQ are highlighted, based on the extensive literature on scaling Gaussian processes and other spline-related models.

Exact inversion can be achieved at low cost through exploiting structure in the kernel matrix. Examples include: the use of kernels with compact support (e.g. Wendland, 2005, Sec. 9) to induce sparsity in the kernel matrix; tensor product kernels (e.g. O'Hagan, 1991) in the context of inverting kernel matrices defined by tensor products of point sets in multivariate problems; using circulant

embedding for a stationary kernel evaluated on an evenly-spaced point set (Davies and Bryant, 2013); and making use of low-rank kernels (e.g. polynomial kernels).

In addition there are many approximate inversion techniques: (i) Reduced rank approximations can reduce the computational cost to $O(nm^2)$ where $m \ll n$ is a parameter controlling the accuracy of the approximation, essentially an effective degree of freedom (Quinero-Candela and Rasmussen, 2005; Bach, 2013; El Alaoui and Mahoney, 2015). (ii) Explicit feature maps designed for additive kernels (Vedaldi and Zisserman, 2012). (iii) Local approximations (Gramacy and Apley, 2015), training only on nearest neighbour data. (iv) Multi-scale approximations, whereby the high-level structure is modelled using a full GP and approximation schemes are applied to lower-level structure (Iske, 2004; Katzfuss, 2015). (v) Fast multipole methods (Wendland, 2005)[Sec.15.1]. (vi) Random approximations of the kernel itself, rather than the kernel matrix, such as random Fourier features (RFF; Rahimi and Recht, 2007), spectral methods (Lazaro-Gredilla et al., 2010; Bach, 2015) and hash kernels (Shi et al., 2009). (RFF have previously been successfully applied in BQ by Briol et al. (2015).) (vii) Parallel programming provides an alternative perspective on complexity reduction, as discussed in (e.g.) Dai et al. (2014).

This does not represent an exhaustive list of the (growing) literature on kernel matrix methods. Efficient use of data structures (Wendland, 2005)[Sec. 14] may also provide significant computational acceleration. Note that the majority of approximate kernel methods do not come with probability models for the additional source of numerical error introduced by the approximation. This could therefore be considered to be sacrificing the philosophical advantages of the probabilistic numerical framework.

C.2 Scalability in Dimension

High-dimensional integrals arising in applications are, in many cases, effectively low-dimensional problems. This can occur either (i) when the distribution Π is effectively concentrated in a low-dimensional manifold in \mathcal{X} (this is responsible for the excellent performance of (MC)MC in certain high-dimensional settings), or (ii) when the integrand f depends on only a subset of its inputs, possibly after a transformation (this is responsible for the excellent performance of QMC methods in certain high-dimensional settings; Dick et al., 2013). The B(MC)MC and BQMC methods that we study provably deliver performance that is at least equivalent to (MC)MC and QMC in settings (i) and (ii) respectively (see Sec. 5.2.3 for an empirical example with $d = 50$). Conversely, when neither Π nor f are effectively low-dimensional, all approaches to integration necessarily suffer from a curse of dimension. For example, for Π uniform on $\mathcal{X} = [0, 1]^d$ and f belonging to a general Sobolev space of order α , no deterministic integration algorithm can exceed the $O(n^{-\alpha/d})$ rate. Clearly this rate becomes arbitrarily slow as d tends to infinity. Nevertheless, we note that BQ estimators remain coherent in an inferential sense, reverting to the prior in this degenerate limit. Indeed, this is natural from a Bayesian point of view since our approximation of the integrand f will become very poor as d grows with n fixed.

We briefly note that a number of alternative approaches exist for problems in which the effective dimensionality is low. In particular, low-dimensional random embeddings project the ambient space into a lower dimensional space using a randomized map, perform computation in that space and then map back the results to the original space (see e.g. Wang et al., 2013, in the context of Bayesian optimisation).

C.3 Regularisation and Noisy Observations

This section studies the impact of numerical regularisation of the kernel matrix on estimator performance. Indeed, it is well known that instability in the numerical inversion of the kernel matrix \mathbf{K} can sometimes lead to a loss of performance (Schaback, 1995). Several techniques have been developed to counteract this issue. Examples include pre-conditioning of the kernel matrix and change of basis tricks (i.e. finding a different kernel spanning the same space of functions but for which the conditioning of the kernel matrix is improved). We refer the interested reader to Wendland (2005)[Chap. 12] for an overview of available methods.

In this section we focus on a simple trick which consists of replacing \mathbf{K} by $\mathbf{K} + \lambda \mathbf{I}$ for some small $\lambda > 0$. Such regularisation can be interpreted in several ways. If added solely to improve numerical stability, $\lambda \mathbf{I}$ is sometimes referred to as *jitter* or a *nugget* term. However, an alternative interpretation, of particular interest here, is that the observed function values f_i are corrupted by noise. In this case the posterior variance is naturally inflated (see e.g. Fig. 10). Below we provide a theoretical study of the convergence and contraction of BMC under the assumption of noisy data, which is equivalent to studying BMC when numerical regularisation is employed.

Consider an homoscedastic Gaussian noise model in which $\mathbf{y} = \mathbf{f} + \mathbf{e}$ is observed, where $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \tau^{-1} \mathbf{I})$. In this case, using the conjugacy of Gaussian variables, it is possible to obtain a closed-form expression for the induced quadrature rule $\hat{\Pi}_{\text{BQ}}^{\mathbf{e}}$ and other quantities of interest by replacing \mathbf{f} by \mathbf{y} and adding a constant term to the diagonal of the kernel matrix of size $\lambda = \tau^{-1}$ (Rasmussen and Williams, 2006). This leads to a probabilistic integrator with

$$\|\hat{\Pi}_{\text{BQ}}^{\mathbf{e}} - \Pi\|_{\mathcal{H}}^2 = \|\hat{\Pi}_{\text{BQ}} - \Pi\|_{\mathcal{H}^*}^2 + \tau^{-1} \|\mathbf{w}^{\text{BQ}}\|_2^2,$$

see (Bach, 2015, Sec. 3.1). Since the term $\|\mathbf{w}^{\text{BQ}}\|_2$ can in general decay more slowly (as $n \rightarrow \infty$) compared to the WCE term $\|\hat{\Pi}_{\text{BQ}} - \Pi\|_{\mathcal{H}^*}$, it comes as no surprise that convergence rates are slower in the noisy data regime, compared to the noiseless regime. This is made precise in the following:

Lemma 6 (BMC with noisy data). *Let $\mathcal{X} = [0, 1]^d$ and consider data \mathbf{y} generated under the homoscedastic Gaussian noise model. Then:*

1. *If \mathcal{H} is an RKHS in the intersection of the Sobolev space \mathcal{H}_α and the Hölder space \mathcal{C}_α for $\alpha > d/2$, we have $\|\hat{\Pi}_{\text{BMC}}^{\mathbf{e}} - \Pi\|_{\mathcal{H}^*} = O_P(n^{-\alpha/(2\alpha+d)})$ and $\mathbb{P}_n[I_D^c] = O_P(\exp(-Cn^{2\alpha/(2\alpha+d)}))$.*
2. *If \mathcal{H} is an RKHS with Gaussian kernel we have $\|\hat{\Pi}_{\text{BMC}}^{\mathbf{e}} - \Pi\|_{\mathcal{H}^*} = O_P(n^{-1/2+\epsilon})$ and $\mathbb{P}_n[I_D^c] = O_P(\exp(-Cn^{1-\epsilon}))$ where $\epsilon > 0$ can be arbitrarily small.*

Proof of Lemma 6. Initially consider fixed states $\{\mathbf{x}_i\}_{i=1}^n$ (i.e. fixing the random seed). Fix a particular integration problem whose true integrand is $f_0 \in \mathcal{H}$. Since the WCE (squared) coincides with the posterior variance, we have from Jensen's inequality

$$\|\hat{\Pi}_{\text{BMC}}^{\mathbf{e}} - \Pi\|_{\mathcal{H}^*}^2 = \mathbb{E}_n[\Pi[f] - \mathbb{E}_n[\Pi[f]]]^2 = \mathbb{E}_n[\Pi[f - \mathbb{E}_n[f]]]^2 \leq \mathbb{E}_n\|f - \mathbb{E}_n[f]\|_2^2.$$

Here $\mathbb{E}_n = \mathbb{E}[\cdot | \{\mathbf{x}_i^{\text{MC}}, y_i\}_{i=1}^n]$ denotes an expectation with respect to the posterior GP (a slight abuse of notation since \mathbb{E}_n is used for a different purpose in the main text). Noting that $\mathbb{E}_n[f]$ is the variational minimiser of the posterior least squares loss, we have $\mathbb{E}_n\|f - \mathbb{E}_n[f]\|_2^2 \leq \mathbb{E}_n\|f - f_0\|_2^2$. Now, taking an expectation \mathbb{E}_X over the states $\{\mathbf{x}_i\}_{i=1}^n$, viewed as independent draws from Π , we have

$$\mathbb{E}_X \|\hat{\Pi}_{\text{BMC}}^{\mathbf{e}} - \Pi\|_{\mathcal{H}^*}^2 \leq \mathbb{E}_X \mathbb{E}_n \|f - f_0\|_2^2. \quad (11)$$

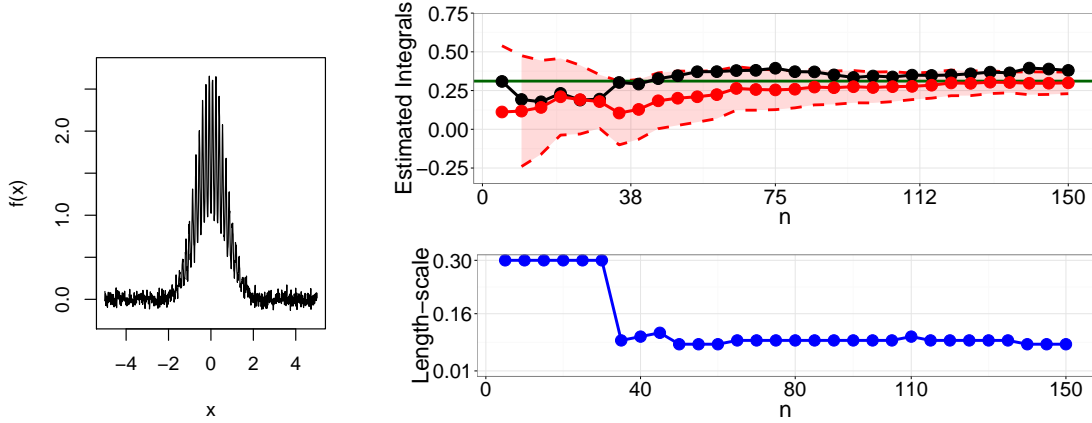


Figure 10: Performance of BMC on test function f_2 with noisy observations. *Left*: Plot of the noisy version of f_2 . *Right*: Estimate of the integral provided by MC (black) and BMC (red) with 95% credible intervals (dotted red lines).

Since the left hand side of Eqn. 9 is independent of f_0 , it suffices to exhibit a particular regression problem f_0 for which the right hand side converges at a known rate. Suppose in addition that $f_0 \in \mathcal{C}_\alpha \cap \mathcal{H}_\alpha$ for $\alpha > d/2$ (this includes for example the function $f_0 \equiv 0$). Then from Theorem 5 of van Der Vaart and van Zanten (2011) we have a scaling relationship $\mathbb{E}_X \mathbb{E}_n \|f - f_0\|_2^2 = O(n^{-2\alpha/(2\alpha+d)})$. From Markov's inequality, convergence in mean implies convergence in probability and thus, combining Eqn. 11 with the scaling relationship, we have $\|\hat{\Pi}_{\text{BMC}}^e - \Pi\|_{\mathcal{H}^*} = O_P(n^{-\alpha/(2\alpha+d)})$.

On the other hand, if we have a Gaussian kernel then we suppose in addition that f_0 is a restriction to $[0, 1]^d$ of an element of $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$, for $r \geq 1$ and $\gamma > 0$, defined to be the set of functions whose Fourier transform $\mathfrak{F}f_0$ satisfies $\int \exp(\gamma\|\xi\|^r) |\mathfrak{F}f_0|^2(\xi) d\xi < \infty$. Again, the function $f_0 \equiv 0$ belongs to $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$. This time, from Theorem 10 of van Der Vaart and van Zanten (2011) we have a scaling relationship $\mathbb{E}_X \mathbb{E}_n \|f - f_0\|_2^2 = O((\log n)^{2/r}/n)$. Since the function $f_0 \equiv 0$ belongs to $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$ for all $r \geq 1$ we conclude, via Markov's inequality as before, that $\|\hat{\Pi}_{\text{BMC}}^e - \Pi\|_{\mathcal{H}^*} = O_P(n^{-1/2+\epsilon})$ where $\epsilon > 0$ can be arbitrarily small. This completes the proof. \square

Thus the effect of measurement noise is to destroy the asymptotic efficiency of BMC over a simple MC estimator; in fact the BMC estimator can become *worse* than the MC estimator in these instances. This should serve as a warning when dealing with kernel matrices that have poor condition number and necessitate careful monitoring of this condition number in computer code.

A brief investigation was undertaken into the issue of noisy data on our test functions from Sec. 5.1. In particular, we consider estimating f_2 with $\mathcal{N}(0, 0.05^2)$ measurement error. As shown on the left-hand side of Fig. 10, the EB procedure manages to account for noise in the observations and adapts uncertainty adequately (here we fixed $\lambda = 1$ and selected σ using EB). In fact, Fig. 11 demonstrates accurate coverage of the credible intervals obtained using EB for this “hard” test function. Thus for small noise (resp. small numerical regularisation) the BQ procedure demonstrated robustness in this example.

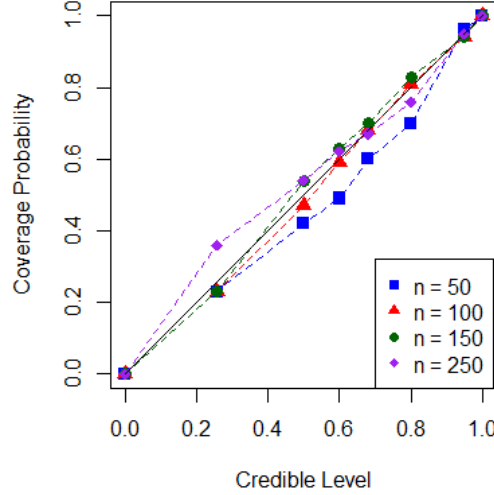


Figure 11: Frequentist coverage of credible intervals for test function f_2 . The empirical Bayes procedure accounts for the observation noise and inflates the credible intervals appropriately in this example.

C.4 Intractable Kernel Means

In this section we propose *approximate Bayesian Quadrature*, ${}_a\hat{\Pi}_{\text{BQ}}$, where the weights ${}_a\mathbf{w}_{\text{BQ}} = \mathbf{K}^{-1}{}_a\Pi[\mathbf{k}(X, \cdot)]$ are an approximation to the optimal BQ weights based on an approximation ${}_a\Pi[\mathbf{k}(X, \cdot)]$ of the kernel mean (see also Prop. 1 in Sommariva and Vianello, 2006). The following lemma demonstrates that we can bound the contribution of this error and inflate our posterior $\mathbb{P}_n \mapsto {}_a\mathbb{P}_n$ to reflect the additional uncertainty due to the approximation, so that uncertainty quantification is still provided. The statistical foundations of such an approach, where updating of belief occurs via an approximation to the likelihood function, are discussed in the recent work of Bissiri et al. (2016).

Lemma 7 (Approximate kernel mean). *Consider an approximation ${}_a\Pi$ to Π of the form ${}_a\Pi = \sum_{j=1}^m {}_aw_j\delta(\cdot - {}_a\mathbf{x}_j)$. Then BQ can be performed analytically with respect to ${}_a\Pi$; denote this estimator by ${}_a\hat{\Pi}_{\text{BQ}}$. Moreover, $\|{}_a\hat{\Pi}_{\text{BQ}} - \Pi\|_{\mathcal{H}^*}^2 \leq \|\hat{\Pi}_{\text{BQ}} - \Pi\|_{\mathcal{H}^*}^2 + \sqrt{n}\|{}_a\Pi - \Pi\|_{\mathcal{H}^*}^2$.*

Proof. Define $\mathbf{z} = \Pi[\mathbf{k}(X, \cdot)]$ and ${}_a\mathbf{z} = {}_a\Pi[\mathbf{k}(X, \cdot)]$. Let $\boldsymbol{\epsilon} = {}_a\mathbf{z} - \mathbf{z}$, write ${}_a\hat{\Pi}_{\text{BQ}} = \sum_{i=1}^n {}_aw_i^{\text{BQ}}\delta(\cdot - \mathbf{x}_i)$ and consider

$$\begin{aligned} \|{}_a\hat{\Pi}_{\text{BQ}} - \Pi\|_{\mathcal{H}^*}^2 &= \|\mu({}_a\hat{\Pi}_{\text{BQ}}) - \mu(\Pi)\|_{\mathcal{H}}^2 \\ &= \left\langle \sum_{i=1}^n {}_aw_i^{\text{BQ}}k(\cdot, \mathbf{x}_i) - \int_{\mathcal{X}} k(\cdot, \mathbf{x})\Pi(d\mathbf{x}), \right. \\ &\quad \left. \sum_{i=1}^n {}_aw_i^{\text{BQ}}k(\cdot, \mathbf{x}_i) - \int_{\mathcal{X}} k(\cdot, \mathbf{x})\Pi(d\mathbf{x}) \right\rangle_{\mathcal{H}} \end{aligned}$$

which simplifies to

$$\begin{aligned}
&= {}_a\mathbf{w}_{\text{BQ}}^T \mathbf{K}_a \mathbf{w}_{\text{BQ}} - 2 {}_a\mathbf{w}_{\text{BQ}}^T \mathbf{z} + \Pi[\mu(\Pi)] \\
&= (\mathbf{K}^{-1} {}_a\mathbf{z})^T \mathbf{K} (\mathbf{K}^{-1} {}_a\mathbf{z}) - 2 (\mathbf{K}^{-1} {}_a\mathbf{z})^T \mathbf{z} + \Pi[\mu(\Pi)] \\
&= (\mathbf{z} + \boldsymbol{\epsilon})^T \mathbf{K}^{-1} (\mathbf{z} + \boldsymbol{\epsilon}) - 2 (\mathbf{z} + \boldsymbol{\epsilon})^T \mathbf{K}^{-1} \mathbf{z} + \Pi[\mu(\Pi)] \\
&= \|\hat{\Pi}_{\text{BQ}} - \Pi\|_{\mathcal{H}^*}^2 + \boldsymbol{\epsilon}^T \mathbf{K}^{-1} \boldsymbol{\epsilon}.
\end{aligned}$$

Use \otimes to denote the tensor product of RKHS. Now, since

$$\epsilon_i = {}_a z_i - z_i = \mu({}_a \hat{\Pi})(\mathbf{x}_i) - \mu(\Pi)(\mathbf{x}_i) = \langle \mu({}_a \hat{\Pi}) - \mu(\Pi), k(\cdot, \mathbf{x}_i) \rangle_{\mathcal{H}},$$

we have:

$$\begin{aligned}
\boldsymbol{\epsilon}^T \mathbf{K}^{-1} \boldsymbol{\epsilon} &= \sum_{i,i'} [\mathbf{K}^{-1}]_{i,i'} \langle \mu({}_a \hat{\Pi}) - \mu(\Pi), k(\cdot, \mathbf{x}_i) \rangle_{\mathcal{H}} \langle \mu({}_a \hat{\Pi}) - \mu(\Pi), k(\cdot, \mathbf{x}_{i'}) \rangle_{\mathcal{H}} \\
&= \left\langle (\mu({}_a \hat{\Pi}) - \mu(\Pi)) \otimes (\mu({}_a \hat{\Pi}) - \mu(\Pi)), \sum_{i,i'} [\mathbf{K}^{-1}]_{i,i'} k(\cdot, \mathbf{x}_i) \otimes k(\cdot, \mathbf{x}_{i'}) \right\rangle_{\mathcal{H} \otimes \mathcal{H}} \\
&\leq \|\mu({}_a \hat{\Pi}) - \mu(\Pi)\|_{\mathcal{H}}^2 \left\| \sum_{i,i'} [\mathbf{K}^{-1}]_{i,i'} k(\cdot, \mathbf{x}_i) \otimes k(\cdot, \mathbf{x}_{i'}) \right\|_{\mathcal{H} \otimes \mathcal{H}}.
\end{aligned}$$

From Prop. 2 we have $\|\mu({}_a \hat{\Pi}) - \mu(\Pi)\|_{\mathcal{H}} = e({}_a \hat{\Pi}; \Pi, \mathcal{H})$ so it remains to show that the second term is equal to \sqrt{n} . Indeed,

$$\begin{aligned}
&\left\| \sum_{i,i'} [\mathbf{K}^{-1}]_{i,i'} k(\cdot, \mathbf{x}_i) \otimes k(\cdot, \mathbf{x}_{i'}) \right\|_{\mathcal{H}}^2 \\
&= \sum_{i,i',l,l'} [\mathbf{K}^{-1}]_{i,i'} [\mathbf{K}^{-1}]_{l,l'} \langle k(\cdot, \mathbf{x}_i) \otimes k(\cdot, \mathbf{x}_{i'}), k(\cdot, \mathbf{x}_l) \otimes k(\cdot, \mathbf{x}_{l'}) \rangle_{\mathcal{H}} \\
&= \sum_{i,i',l,l'} [\mathbf{K}^{-1}]_{i,i'} [\mathbf{K}^{-1}]_{l,l'} [\mathbf{K}]_{il} [\mathbf{K}]_{i'l'} = \text{tr}[\mathbf{K} \mathbf{K}^{-1} \mathbf{K} \mathbf{K}^{-1}] = n.
\end{aligned}$$

This completes the proof. \square

Under this method, the posterior variance ${}_a \mathbb{V}_n[\Pi[f]] := \|{}_a \hat{\Pi}_{\text{BQ}} - \Pi\|_{\mathcal{H}^*}^2$ cannot be computed in closed-form, but computable upper-bounds can be obtained and these can then be used to propagate numerical uncertainty through the remainder of our statistical task. The idea here is to make use of the triangle inequality:

$$\|{}_a \hat{\Pi}_{\text{BQ}} - \Pi\|_{\mathcal{H}^*} \leq \|{}_a \hat{\Pi}_{\text{BQ}} - {}_a \Pi\|_{\mathcal{H}^*} + \|{}_a \Pi - \Pi\|_{\mathcal{H}^*}. \quad (12)$$

The first term on the RHS is now available analytically; from Prop. 1 its square is ${}_a \Pi {}_a \Pi[k(\cdot, \cdot)] - {}_a \Pi[\mathbf{k}(\cdot, X)] \mathbf{K}^{-1} {}_a \Pi[\mathbf{k}(X, \cdot)]$. For the second term, explicit upper bounds exist in the case where states ${}_a \mathbf{x}_i$ are independent random samples from Π . For instance, from (Song, 2008, Thm. 27) we have, for a radial kernel k , uniform ${}_a w_j = m^{-1}$ and independent ${}_a \mathbf{x}_i \sim \Pi$,

$$\|{}_a \Pi - \Pi\|_{\mathcal{H}^*} \leq \frac{2}{\sqrt{m}} \sup_{\mathbf{x}} \sqrt{k(\mathbf{x}, \mathbf{x})} + \sqrt{\frac{\log(2/\delta)}{2m}} \quad (13)$$

with probability at least $1 - \delta$. (For dependent ${}_a\mathbf{x}_j$, the m in Eqn. 13 can be replaced with an estimate for the effective sample size.) Write $C_{n,\alpha,\delta}$ for a $100(1 - \alpha)\%$ credible interval for $\Pi[f]$ defined by the conservative upper bound described in Eqns. 12 and 13. Then we conclude that $C_{n,\alpha,\delta}$ is $100(1 - \alpha)\%$ credible interval with probability at least $1 - \delta$.

Note that, even though the credible region has been inflated, it still contracts to the truth, since the first term on the RHS in Lemma 7 can be bounded by the sum of $\|{}_a\hat{\Pi}_{\text{BQ}} - \Pi\|_{\mathcal{H}^*}$ and $\|{}_a\Pi - \Pi\|_{\mathcal{H}^*}$, both of which vanish as $n, m \rightarrow \infty$. The resulting (conservative) posterior ${}_a\mathbb{P}_n$ can be viewed as a updating of beliefs based on an approximation to the likelihood function; the statistical foundations of such an approach are made clear in the recent work of Bissiri et al. (2016).

We pause to briefly discuss the utility and significance of such an approach. Obviously, the new approximation problem (that of approximating Π with ${}_a\Pi$) could also be computed with a BQ method, and we may hence end up in an “infinite regress” scenario (O’Hagan, 1991), where the new kernel mean is itself unknown and so on. However, one level of approximation may be enough in many scenarios. Indeed, by using standard Monte Carlo to select $\{{}_a\mathbf{x}_j\}_{j=1}^m$ and increasing m sufficiently faster than n , the error term $\sqrt{n}\|{}_a\Pi - \Pi\|_{\mathcal{H}^*}^2$ can be made to vanish faster than $\|\hat{\Pi}_{\text{BQ}} - \Pi\|_{\mathcal{H}^*}^2$ and hence the WCE for ${}_a\hat{\Pi}_{\text{BQ}}$ will be asymptotically identical to the WCE for the (intractable) exact BQ estimator $\hat{\Pi}_{\text{BQ}}$. Therefore, it will be reasonable to expend computational effort on raising m in settings where evaluation of the integrand constitutes the principal computational. This is because approximating the kernel mean only requires sampling m times, but does not require us to evaluate the integrand.

D Additional Numerical Results

This section presents additional numerical results concerning both the calibration of uncertainty for multiple parameters and in higher dimensions (Sec. D.1) and rates of posterior contraction (Sec. D.2).

D.1 Empirical Bayes Calibration

Calibration in 1D In Fig. 12 (top row) we study the quantification of uncertainty provided by EB in the same setup as in the main text, but optimizing over both length-scale parameter σ and magnitude parameter λ . For both “easy” and “hard” test functions, we notice that EB led to over-confident inferences in the “low n ” regime, but attains approximately correct frequentist coverage for larger n .

Calibration in 5D The experiments of Sec. 5.1, based on BMC, were repeated with the same test functions f_1 and f_2 , with various values of kernel parameter $p = \alpha + 1/2 \in \{3/2, 5/2, 7/2\}$ in dimension $d = 5$. Results are shown in Fig. 12 (bottom row). Clearly more integrand evaluations are required for EB to attain a good frequentist coverage of the credible intervals, due to the curse of dimension. However, the frequentist coverage was reasonable for large enough n in this example.

D.2 Empirical Rates of Posterior Contraction

Here we attempt to validate the theoretical results presented in Sec. 3 for the rates of posterior contraction for BMC and BQMC.

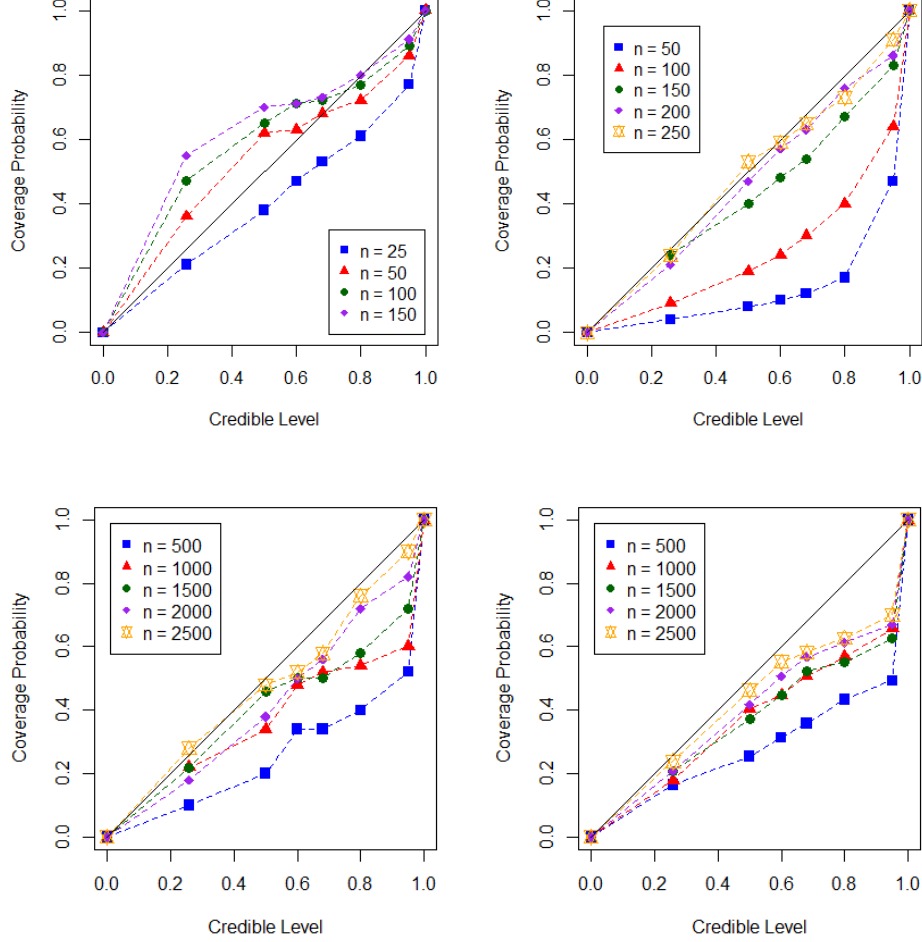


Figure 12: Evaluation of uncertainty quantification provided by EB for the length-scale σ and magnitude λ in $d = 1$ (top plots) or for σ only in $d = 5$ (bottom plots). Coverage frequencies $C_{n,\alpha}$ (computed from 100 (top) or 50 (bottom) realisations) were compared against notional $100(1-\alpha)\%$ Bayesian credible regions for varying level α . *Left*: “easy” function f_1 . *Right*: “hard” test function f_2 .

BMC In Sec. 3.2.1, it was proven that the BMC standard deviation $\|\hat{\Pi}_{\text{BMC}} - \Pi\|_{\mathcal{H}^*} = O_P(n^{-\alpha/d+\epsilon})$ when \mathcal{H} is a Sobolev space of order $\alpha > d/2$. Fig. 13 (top row) depicts empirical convergence results obtained for $d = 1$ (left) and $d = 5$ (right), for *one* typical realisation. In the one dimensional case, the $O_P(n^{-\alpha/d+\epsilon})$ theoretical convergence rates are broadly attained (and indeed exceeded by at most one Monte Carlo, as explained in the main text). At larger values of n , numerical regularisation takes effect, as predicted by the analysis in Supplement C.3. In the higher dimensional case, the only rate proven in this paper is $O_P(n^{-1/2})$ since $\alpha < d/2$ in all cases $p = \alpha + 1/2 \in \{3/2, 5/2, 7/2\}$ considered. These results offer broad agreement with our theoretical results.

BQMC In Sec. 3.2.2 it was proven that the BQMC standard deviation $\|\hat{\Pi}_{\text{BQMC}} - \Pi\|_{\mathcal{H}^*} = O(n^{-\alpha+\epsilon})$ when \mathcal{H} is a Sobolev space of dominating mixed smoothness and order $\alpha > 1/2$. Fig. 13

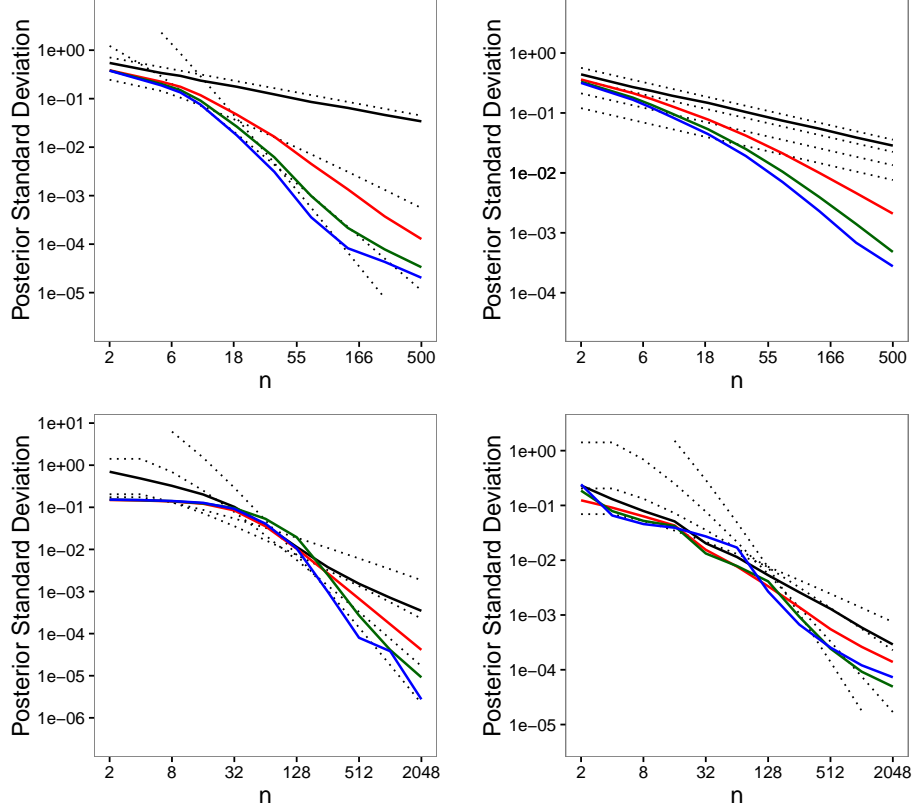


Figure 13: Posterior standard deviation for *one realisation* of BMC and BQMC on $[0, 1]^d$ for $d = 1$ (left) and $d = 5$ (right). Here we considered BMC in Sobolev spaces \mathcal{H}_α (top row), and BQMC in Sobolev spaces of mixed dominating smoothness \mathcal{S}_α (bottom row). The results are obtained using tensor product Matérn kernels of smoothness $\alpha = 3/2$ (red), $\alpha = 5/2$ (green) and $\alpha = 7/2$ (blue). Dotted lines represent the theoretical convergence rates established for each kernel. The black line represents the corresponding standard Monte Carlo or quasi-Monte Carlo rate. Kernel parameters were fixed to $(\sigma, \lambda) = (0.02, 1)$ (top left), $(\sigma, \lambda) = (1.2, 1)$ (top right), $(\sigma, \lambda) = (0.005, 1)$ (bottom left) and $(\sigma, \lambda) = (1, 0.5)$ (bottom right).

(bottom row) depicts empirical convergence results obtained for $d = 1$ (left) and $d = 5$ (right), for *one* typical realisation. In the one dimensional case, the $O(n^{-\alpha+\epsilon})$ theoretical convergence rate is broadly attained in all cases $p = \alpha + 1/2 \in \{3/2, 5/2, 7/2\}$ considered. However, in the $d = 5$ case, the rates are not observed for the number n of evaluations considered. This helps us demonstrate the important point that the rates we provide are asymptotic, and may require large values of n before being observed in practice.

E Supplemental Information for Case Studies

E.1 Case Study #1

MCMC Established MCMC techniques are available for obtaining samples from power posteriors. In this paper we used the manifold Metropolis-adjusted Langevin algorithm (Girolami and Calderhead, 2011) in combination with population MCMC, a combination also explored in Oates et al. (2016c). Population MCMC shares information across temperatures during sampling, yet previous work has not leveraged evaluation of the log-likelihood f from one sub-chain t_i to inform estimates derived from other sub-chains $t_{i'}$, $i' \neq i$. In contrast, this occurs naturally in the probabilistic integration framework, as described in the main text.

Here MCMC was used to generate a small number, $n = 200$, of samples on a per-model basis, in order to simulate a scenario where numerical error in computation of marginal likelihood will be non-negligible. A temperature ladder with $m = 10$ rungs was employed, for the same reason, according to the recommendation of Calderhead and Girolami (2009). No convergence issues were experienced; the same MCMC set-up has previously been successfully used in Oates et al. (2016c).

Prior Elicitation Here we motivate a prior for the unknown function g based on the work of Calderhead and Girolami (2009), who advocated the use of a power-law schedule $t_i = (\frac{i-1}{m-1})^5$, $i = 1, \dots, m$, based on an extensive empirical comparison of possible schedules. A “good” temperature schedule approximately satisfies the criterion

$$|g(t_i)(t_{i+1} - t_i)| \approx m^{-1},$$

on the basis that this allocates equal area to the portions of the curve g that lie between t_i and t_{i+1} , controlling bias for the trapezium rule. Substituting $t_i = (\frac{i-1}{m-1})^5$ into this optimality criterion produces

$$|g(t_i)|((i+1)^5 - i^5) \approx m^4.$$

Now, letting $i = \theta m$

$$|g(\theta^5)| (5\theta^4 m^4 + o(m^4)) \approx m^4.$$

Formally treating θ as continuous and taking the $m \rightarrow \infty$ limit produces

$$|g(\theta^5)| \approx \frac{1}{5\theta^4} \implies |g(t)| \approx \frac{1}{5t^{4/5}}.$$

From this we conclude that the transformed function $h(t) = 5t^{4/5}g(t)$ is approximately stationary and can reasonably be assigned a stationary GP prior. However, in an importance sampling transformation we require that $\pi(t)$ has support over $[0, 1]$. For this reason we took

$$\pi(t) = \frac{1.306}{0.01 + 5t^{4/5}}$$

in our experiments.

Variance Computation The covariance matrix Σ cannot be obtained in closed-form due to intractability of the kernel mean $\Pi_{t_i}[k_f(\cdot, \theta)]$. We therefore explored an approximation ${}_a\Sigma$ such that plugging in ${}_a\Sigma$ in place of Σ provides an approximation to the posterior variance $\mathbb{V}_n[\log p(\mathbf{y})]$ for the log-marginal likelihood. This took the form

$${}_a\Sigma_{i,j} := {}_a\Pi_{t_i}{}_a\Pi_{t_j}[k_f(\cdot, \cdot)] - {}_a\Pi_{t_i}[\mathbf{k}_f(\cdot, X)]\mathbf{K}_f^{-1}{}_a\Pi_{t_j}[\mathbf{k}_f(X, \cdot)]$$

where an empirical distribution ${}_a\Pi = \frac{1}{100} \sum_{i=1}^{100} \delta(\cdot - \mathbf{x}_i)$ was employed based on the first $m = 100$ samples, while the remaining samples $X = \{\mathbf{x}_i\}_{i=101}^{200}$ were reserved for the kernel computation. This heuristic approach becomes exact as $m \rightarrow \infty$, in the sense that ${}_a\Sigma_{i,j} \rightarrow \Sigma_{i,j}$, but underestimates covariance at finite values of m . Results in the main text suggest that the heuristic performs reasonably in practice, in the sense of providing reasonable uncertainty quantification.

We note that the alternative gradient-based kernel construction of Oates et al. (2016a) would permit *exact* formulae for kernel means here, with no need to employ heuristics as above. Our motivation for proceeding with a generic kernel k_f here was not to complicate the presentation by introducing a non-trivial kernel construction.

Kernel Choice In experiments below, both k_f and k_h were taken to be Gaussian covariance functions; for example: $k_f(\mathbf{x}, \mathbf{x}') = \lambda_f \exp(-\|\mathbf{x} - \mathbf{x}'\|_2^2 / 2\sigma_f^2)$ parametrised by λ_f and σ_f . This choice was made to capture infinite differentiability of both integrands f and h involved in the example below. For this application we found that, while the σ parameters were possible to learn from data using EB, the λ parameters required a large number of data to pin down. Therefore, for these experiments we fixed $\lambda_f = 0.1 \times \text{mean}(f_{i,j})$ and $\lambda_h = 0.01 \times \text{mean}(h_i)$. In both cases the remaining kernel parameters σ were selected using EB. Results in Fig. 14 present the posterior estimates for the marginal likelihoods $p_i = p(\mathbf{y}|\mathcal{M}_i)$ that take into account numerical error, contrasted with the single point estimates obtained by ignoring numerical error (the standard approach).

Data Generation As a test-bed that captures the salient properties of model selection discussed in the main text, we considered variable selection for logistic regression:

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\beta}) &= \prod_{i=1}^N p_i(\boldsymbol{\beta})^{y_i} [1 - p_i(\boldsymbol{\beta})]^{1-y_i} \\ \text{logit}(p_i(\boldsymbol{\beta})) &= \gamma_1 \beta_1 x_{i,1} + \dots \gamma_d \beta_d x_{i,d}, \quad \gamma_1, \dots, \gamma_d \in \{0, 1\} \end{aligned}$$

where the model \mathcal{M}_k specifies the active variables via the binary vector $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_d)$. A model prior $p(\boldsymbol{\gamma}) \propto d^{-\|\boldsymbol{\gamma}\|_1}$ was employed. Given a model \mathcal{M}_k , the active parameters β_j were endowed with independent priors $\beta_j \sim \mathcal{N}(0, \tau^{-1})$, where here $\tau = 0.01$.

A single dataset of size $N = 200$ were generated from model \mathcal{M}_1 with parameter $\boldsymbol{\beta} = (1, 0, \dots, 0)$; as such the problem is under-determined (there are in principle $2^{10} = 1024$ different models) and the true model is not well-identified. The selected model is thus sensitive to numerical error in the computation of marginal likelihood. In practice we limited the model space to consider only models with $\sum \gamma_i \leq 2$; this speeds up the computation and, in this particular case, only rules out models that have much lower posterior probability than the actual MAP model. There were thus 56 models under comparison.

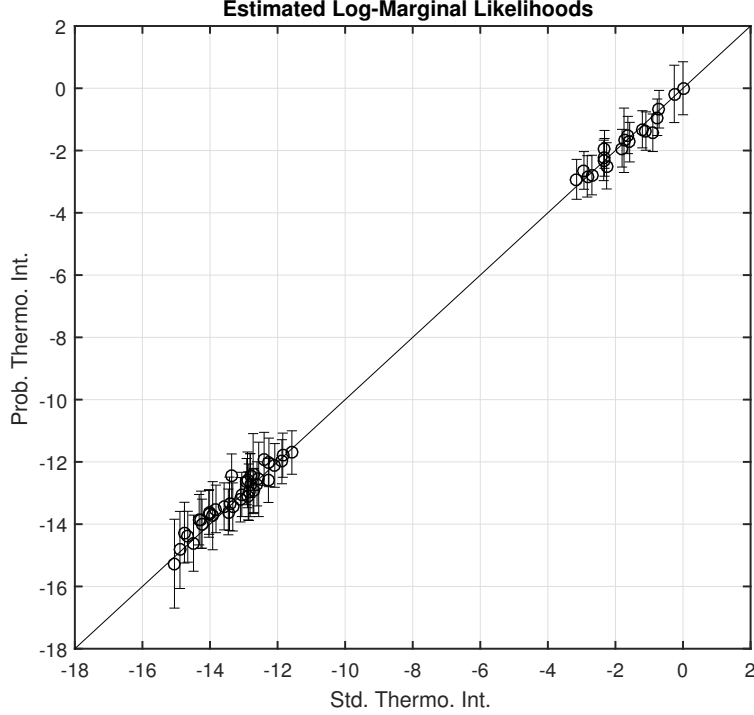


Figure 14: Probabilistic thermodynamic integration; estimates of marginal likelihoods $p_i = p(\mathbf{y}|\mathcal{M}_i)$. On the x-axis we show point estimates obtained by ignoring numerical error (the standard approach). On the y-axis we present the posterior mean estimates and \pm one posterior standard deviation that aims to capture the extent of numerical error.

E.2 Case Study #2

Background on the model The Teal South model is a PDE computer model for an oil reservoir. The model studied is on an 11×11 grid with 5 layers. It has 9 parameters representing physical quantities of interest, for which we would like to do inference using MCMC. These include horizontal permeabilities for each of the 5 layers, the vertical to horizontal permeability ratio, aquifer strength, rock compressibility and porosity. For our experiments, we used a Gaussian process approximation of the likelihood model documented in Lan et al. (2016) in order to speed up MCMC; however this might be undesirable in general due to the additional uncertainty associated with the approximation in the results obtained.

Kernel Choice The numerical results in Sec. 5.2.2 were obtained using a Matérn $\alpha = 3/2$ kernel given by $k(r) = \lambda^2(1 + \sqrt{3}r/\sigma) \exp(-\sqrt{3}r/\sigma)$ where $r = \|\mathbf{x} - \mathbf{y}\|_2$, which corresponds to the Sobolev space $\mathcal{H}_{3/2}$. We note that $f \in \mathcal{H}_{3/2}$ since we know that f has infinitely many derivatives in this application. We used EB over the length-scale parameter σ , but fixed the magnitude parameter to $\lambda = 1$ (we once again found that λ required a large number of data to learn accurately).

Variance Computation Due to intractability of the posterior distribution, the kernel mean $\mu(\pi)$ is unavailable in closed form. To overcome this, the methodology in Sec. 4.2 was employed to obtain an empirical estimate of the kernel mean (half of the MCMC samples were used with BQ

weights to approximate the integral and the other half with MC weights to approximate the kernel mean). The framework described in Supplement C.4, and in particular of Eqn. 12, was used to upper bound the intractable BQ posterior variance. For the upper bound to hold, states \mathbf{x}_j must be independent samples from Π , whereas here they were obtained using MCMC and were therefore not independent. In order to ensure that MCMC samples were “as independent as possible” we employed sophisticated MCMC methodology developed by Lan et al. (2016) that provides low autocorrelation. Nevertheless, we emphasise that there is a gap between theory and practice here that we hope to fill in future research. For the results in this paper we fixed $\delta = 0.05$ in Eqn. 13, so that $C_{n,\alpha} = C_{n,\alpha,0.05}$ is essentially a $95(1 - \alpha)\%$ credible interval. A formal investigation into the theoretical properties of the uncertainty quantification studied by these methods is not provided in this paper.

E.3 Case Study #3

Kernel Choice The (canonical) *weighted* Sobolev space $\mathcal{S}_{\alpha,\gamma}$ is defined by taking each of the component spaces \mathcal{H}_u to be Sobolev spaces of dominating mixed smoothness \mathcal{S}_α . i.e. the space \mathcal{H}_u is norm-equivalent to a tensor product of $|u|$ one-dimensional Sobolev spaces, each with smoothness parameter α . Constructed in this way, $\mathcal{S}_{\alpha,\gamma}$ is an RKHS with kernel

$$k_{\alpha,\gamma}(\mathbf{x}, \mathbf{x}') = \sum_{u \subseteq \mathcal{I}} \gamma_u \prod_{i \in u} \left(\sum_{k=1}^{\alpha} \frac{B_k(x_i)B_k(x'_i)}{(k!)^2} - (-1)^\alpha \frac{B_{2\alpha}(|x_i - x'_i|)}{(2\alpha)!} \right),$$

where the B_k are Bernoulli polynomials. For example, taking $\alpha = 1$ we have the kernel

$$k_{1,\gamma}(\mathbf{x}, \mathbf{x}') = \sum_{u \subseteq \mathcal{I}} \gamma_u \prod_{i \in u} \left(\frac{x_i^2}{2} + \frac{(x'_i)^2}{2} - \frac{x_i}{2} - \frac{x'_i}{2} - \frac{|x_i - x'_i|}{2} + \frac{1}{3} \right),$$

and tractable kernel mean $\mu(\Pi)(\mathbf{x}) = \int_{\mathcal{X}} k_{1,\gamma}(\mathbf{x}, \mathbf{x}') d\mathbf{x}' = \gamma_\emptyset$.

Theoretical Results In finite dimensions $d < \infty$, we can construct a higher-order digital net that attains optimal QMC rates for weighted Sobolev spaces:

Theorem 4. *Let \mathcal{H} be an RKHS that is norm-equivalent to $\mathcal{S}_{\alpha,\gamma}$. Then BQMC based on a digital $(t, \alpha, 1, \alpha m \times m, d)$ -net over \mathbb{Z}_b attains the optimal rate $\|\hat{\Pi}_{BQMC} - \Pi\|_{\mathcal{H}^*} = O(n^{-\alpha+\epsilon})$ for any $\epsilon > 0$, where $n = b^m$. Hence $\mathbb{P}_n[I_D^c] = O(\exp(-Cn^{2\alpha-\epsilon}))$.*

Proof. This follows by combining Thm. 15.21 of Dick and Pillichshammer (2010) with Lemma 2. \square

The QMC rules in Theorem 4 do not explicitly take into account the values of the weights γ . An algorithm that tailors QMC states to specific weights γ is known as the *component by component* (CBC) algorithm; further details can be found in (Kuo, 2003). In principle the CBC algorithm can lead to improved rate constants in high dimensions, because effort is not wasted in directions where f varies little, but the computational overheads are also greater. We did not consider CBC algorithms for BQMC in this paper.

Note that the weighted Hilbert space framework allows us to bound the WCE *independently of dimension* providing that $\sum_{u \subseteq \mathcal{I}} \gamma_u < \infty$ (Sloan and Woźniakowski, 1998). This justifies the use of

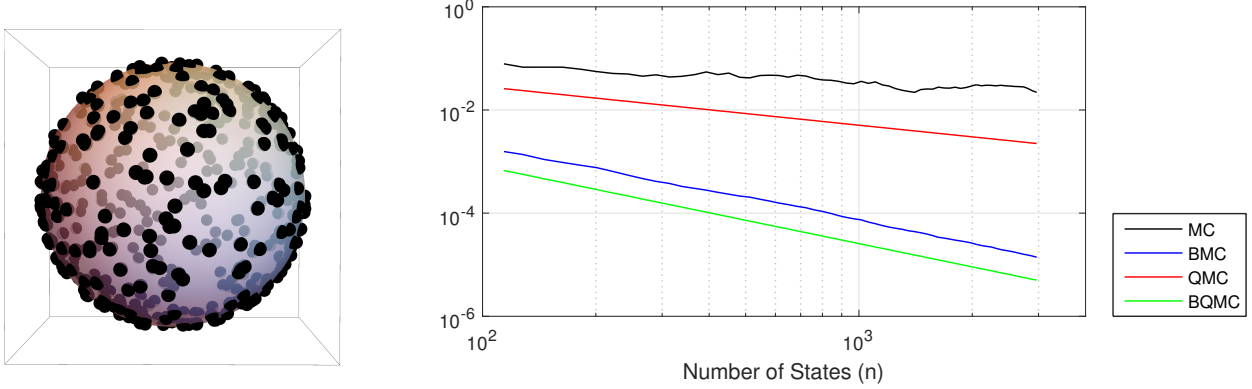


Figure 15: Application to global illumination integrals in computer graphics. *Left:* A spherical t -design over \mathbb{S}^2 . *Right:* The WCE, or worst-case-error, for Monte Carlo (MC), Bayesian MC (BMC), Quasi MC (QMC) and Bayesian QMC (BQMC).

“high-dimensional” in this context; the posterior variance can be bounded independently of dimension for these RKHSs. Analogous results for functional approximation were provided by Fasshauer et al. (2012) for the Gaussian kernel. Further details are provided in Sec. 4.1 of (Dick et al., 2013).

E.4 Case Study #4

Kernel Choice The function spaces that we consider are Sobolev-like spaces $\mathcal{H}_\alpha(\mathbb{S}^d)$ for $\alpha > d/2$, obtained using the reproducing kernel $k(\mathbf{x}, \mathbf{x}') = \sum_{l=0}^{\infty} \lambda_l P_l^{(d)}(\mathbf{x} \cdot \mathbf{x}')$, $\mathbf{x}, \mathbf{x}' \in \mathbb{S}^d$, where $\lambda_l \asymp (1+l)^{-2\alpha}$ and $P_l^{(d)}$ are normalised Gegenbauer polynomials (for $d = 2$ these are also known as Legendre polynomials) (Brauchart et al., 2014). A particularly simple expression for the kernel in $d = 2$ and Sobolev-like space $\alpha = 3/2$ can be obtained by taking $\lambda_0 = 4/3$ along with $\lambda_l = -\lambda_0 \times (-1/2)_l / (3/2)_l$ where $(a)_l = a(a+1)\dots(a+l-1) = \Gamma(a+l)/\Gamma(a)$ is the Pochhammer symbol. Specifically, these choices produce

$$k(\mathbf{x}, \mathbf{x}') = \frac{8}{3} - \|\mathbf{x} - \mathbf{x}'\|_2, \quad \mathbf{x}, \mathbf{x}' \in \mathbb{S}^2.$$

This kernel is associated with a tractable kernel mean $\mu(\Pi)(\mathbf{x}) = \int_{\mathbb{S}^2} k(\mathbf{x}, \mathbf{x}') \Pi(d\mathbf{x}') = 4/3$ and hence the initial error is also available $\Pi[\mu(\Pi)] = \int_{\mathbb{S}^2} \mu(\Pi)(\mathbf{x}) \Pi(d\mathbf{x}') = 4/3$.

Theoretical Results The states $\{\mathbf{x}_i\}_{i=1}^n$ could be generated as MC samples. In that case, analogous results to those obtained in Sec. 3.2.1 can be obtained. Specifically, from Thm. 7 of Brauchart et al. (2014) and Bayesian re-weighting (Lemma 2), classical MC leads to slow convergence $\|\hat{\Pi}_{\text{MC}} - \Pi\|_{\mathcal{H}^*} = O_P(n^{-1/2})$. The regression bound argument (Lemma 3) together with a functional approximation result in Le Gia et al. (2012, Thm. 3.2), gives a faster rate for BMC of $\|\hat{\Pi}_{\text{BMC}} - \Pi\|_{\mathcal{H}^*} = O_P(n^{-3/4})$ in dimension $d = 2$. (For brevity the details are omitted.)

Rather than focus on MC methods, we present results based on spherical QMC point sets. We briefly introduce the concept of a *spherical t -design* (Bondarenko et al., 2013) which is defined as a set $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{S}^d$ satisfying $\int_{\mathbb{S}^d} f d\Pi = \frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i)$ for all polynomials $f : \mathbb{S}^d \rightarrow \mathbb{R}$ of degree at most t . (i.e. f is the restriction to \mathbb{S}^d of a polynomial in the usual Euclidean sense $\mathbb{R}^{d+1} \rightarrow \mathbb{R}$).

Theorem 5. *For all $d \geq 2$ there exists C_d such that for all $n \geq C_d t^d$ there exists a spherical t -design on \mathbb{S}^d with n states. Moreover, for $\alpha = 3/2$ and $d = 2$, the use of a spherical t -designs leads to a rate $\|\hat{\Pi}_{BQMC} - \Pi\|_{\mathcal{H}^*} = O(n^{-3/4})$ and $\mathbb{P}_n[I_D^c] = O(\exp(-Cn^{3/2}))$.*

Proof. This property of spherical t -designs follows from combining Hesse and Sloan (2005); Bondarenko et al. (2013) and Lemma 2. \square

The rate in Thm. 5 is best-possible in the space $\mathcal{H}_{3/2}(\mathbb{S}^2)$ (Brauchart et al., 2014) and, unlike the result for BMC, is fully deterministic. (Empirical evidence in Marques et al. (2015) suggests that BQMC attains faster rates than BMC in RKHS that are smoother than $\mathcal{H}_{3/2}(\mathbb{S}^2)$.) Although explicit spherical t -designs are not currently known in closed-form, approximately optimal point sets have been computed numerically to high accuracy. Our experiments were based on such point sets provided by R. Womersley on his website <http://web.maths.unsw.edu.au/~rsw/Sphere/EffSphDes/sf.html> [Accessed 24 Nov. 2015].

Additional theoretical results on point estimates provided by kernel quadrature can be found in Fuselier et al. (2014). In particular they consider stability of the associated linear systems that must be solved to obtain BQ weights. Some alternative kernels for spherical integration are considered in Fuselier et. al. (2013).

Numerical Results The environment map used in this example is freely available at: <http://www.hdrilabs.com/sibl/archive.html>.

In Fig. 15, the value of the WCE is plotted for each of the four methods considered (MC, QMC, BMC, BQMC) as the number of states increases. Both BMC and BQMC appear to attain the same rate for $\mathcal{H}_{3/2}(\mathbb{S}^2)$, although BQMC provides a constant factor improvement over BMC. Note that $O(n^{-3/4})$ was shown by Brauchart et al. (2014) to be best-possible for a deterministic method in the space $\mathcal{H}_{3/2}(\mathbb{S}^2)$.

References

- Bach, F. (2013). Sharp analysis of low-rank kernel matrix approximations. In *Int. Conf. Learn. Theory*, 185–209.
- Dai, B., Xie, B., He, N., Liang, Y., Raj, A., Balcan, M., and Song, L. (2014). Scalable kernel methods via doubly stochastic gradients. In *Adv. Neur. Inf. Proc. Sys.*, 3041–3049.
- Davies, T. M. and Bryant, D. (2013). On circulant embedding for Gaussian random fields in R. *J. Stat. Softw.*, 55(9):1–21.
- El Alaoui, A. and Mahoney, M. W. (2015). Fast randomized kernel methods with statistical guarantees. In *Adv. Neur. Inf. Proc. Sys.*, 775–783.
- Fasshauer, G. E., Hickernell, F. J. and Woźniakowski, H. (2012). On dimension-independent rates of convergence for function approximation with Gaussian kernels. *SIAM J. Numer. Anal.*, 50(1): 247–271.
- Fuselier, E., Hangelbroek, T., Narcowich, F. J., Ward, J. D. and Wright, G. B. (2013). Localized bases for kernel spaces on the unit sphere. *SIAM J. Numer. Anal.*, 51(5):2538–2562.

- Fuselier, E., Hangelbroek, T., Narcowich, F. J., Ward, J. D. and Wright, G. B. (2014). Kernel based quadrature on spheres and other homogeneous spaces. *Numer. Math.*, 127(1):57–92.
- Gramacy, R. B. and Apley, D. W. (2015). Local Gaussian process approximation for large computer experiments. *J. Comput. Graph. Statist.*, 24(2):561–578.
- Hesse, K. and Sloan, I. A. (2005). Worst-case errors in a Sobolev space setting for cubature over the sphere S². *Bull. Aust. Math. Soc.*, 71(1):81–105.
- Iske, A. (2004). *Multiresolution methods in scattered data modelling*. Springer Science & Business Media.
- Katzfuss, M. (2015). A multi-resolution approximation for massive spatial datasets. *arXiv:1507.04789*.
- Kuo, F. Y. (2003). Component-by-component constructions achieve the optimal rate of convergence for multivariate integration in weighted Korobov and Sobolev spaces. *J. Complexity*, 19(3):301–320.
- Lazaro-Gredilla, M., Quinero-Candela, J., Rasmussen, C. E. and Figueiras-Vidal, A. R. (2010). Sparse spectrum Gaussian process regression. *J. Mach. Learn. Res.*, 11:1865–1881.
- Le Gia, Q. T., Sloan, I. H. and Wendland, H. (2012). Multiscale approximation for functions in arbitrary Sobolev spaces by scaled radial basis functions on the unit sphere. *Appl. Comput. Harmon. Anal.*, 32:401–412.
- Quinero-Candela, J. and Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *J. Mach. Learn. Res.*, 6:1939–1959.
- Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. In *Adv. Neur. Inf. Proc. Sys.*, 1177–1184.
- Shi, Q., Petterson, J., Dror, G., Langford, J., Smola, A. and Vishwanathan, S. V. N. (2009). Hash kernels for structured data. *J. Mach. Learn. Res.*, 10:2615–2637.
- Sloan, I. H. and Woźniakowski, H. (1998). When are quasi-Monte Carlo algorithms efficient for high dimensional integrals? *J. Complexity*, 14(1):1–33.
- Sommariva, A. and Vianello, M. (2006). Numerical cubature on scattered data by radial basis functions. *Computing*, 76(3-4):295–310.
- Tsybakov, A. B. (2008). *Introduction to nonparametric estimation*. Springer Science & Business Media.
- Vedaldi, A. and Zisserman, A. (2012). Efficient additive kernels via explicit feature maps. In *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(3):480–492.
- Wang, Z., Zoghi, M., Hutter, F., Matheson, D. and de Freitas, N. (2013). Bayesian optimization in high dimensions via random embeddings. In *Proc. Int. Joint Conf. Artif. Intell.*, 1778–1784.
- Wu, Z. and Schaback, R. (1993). Local error estimates for radial basis function interpolation of scattered data. *IMA J. Numer. Anal.*, 13(1):13–27.